

Artificial intelligence in medical image analysis

Citation for published version (APA):

Ibrahim, A. K. (2022). *Artificial intelligence in medical image analysis: robustness and applications*. [Doctoral Thesis, Maastricht University, Université de Liège]. ProefschriftMaken. <https://doi.org/10.26481/dis.20220330ai>

Document status and date:

Published: 01/01/2022

DOI:

[10.26481/dis.20220330ai](https://doi.org/10.26481/dis.20220330ai)

Document Version:

Publisher's PDF, also known as Version of record

Please check the document version of this publication:

- A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
- The final author version and the galley proof are versions of the publication after peer review.
- The final published version features the final layout of the paper including the volume, issue and page numbers.

[Link to publication](#)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.umlib.nl/taverne-license

Take down policy

If you believe that this document breaches copyright please contact us at:

repository@maastrichtuniversity.nl

providing details and we will investigate your claim.

Artificial intelligence in medical image analysis: robustness and applications

Abdalla Ibrahim

Cover: Turkey Refaee
Layout: Dennis Hendriks || ProefschriftMaken.nl
Printed by: ProefschriftMaken.nl

ISBN: 978-94-6423-652-1

Copyright © 2022, Abdalla Ibrahim

This thesis was accomplished with financial support from the Liege-Maastricht Imaging Valley Grant, project “DEEP-NUCLE”

Artificial intelligence in medical image analysis: robustness and applications

DISSERTATION

to obtain the degree of Doctor
at Maastricht University

by the authority of the rector magnificus Prof.dr. Pamela Habibović
and

to obtain the degree of Docteur en sciences médicales (Doctor in Medical Sciences)
at the Université de Liège

by the authority of the rector magnificus Prof. dr. Pierre Wolper

in accordance with the decision of the Board of Deans,
to be defended in public on

Wednesday 30 March 2022 at 16.00 hours

by

Abdalla Khalil Ibrahim

born on 7th of January 1991
in Medani, Sudan

Promoters:

Prof. dr. Philippe Lambin

Prof. dr. Felix M. Mottaghy

Prof. dr. Roland Hustinx (University of Liège)

Co-promoter:

Dr. Andrew D.A Maidment (University of Pennsylvania)

Thesis Assessment Committee:

Prof. Dr. Manon van Engeland (Chair)

Prof. Dr. Lawrence Schwartz (Columbia University)

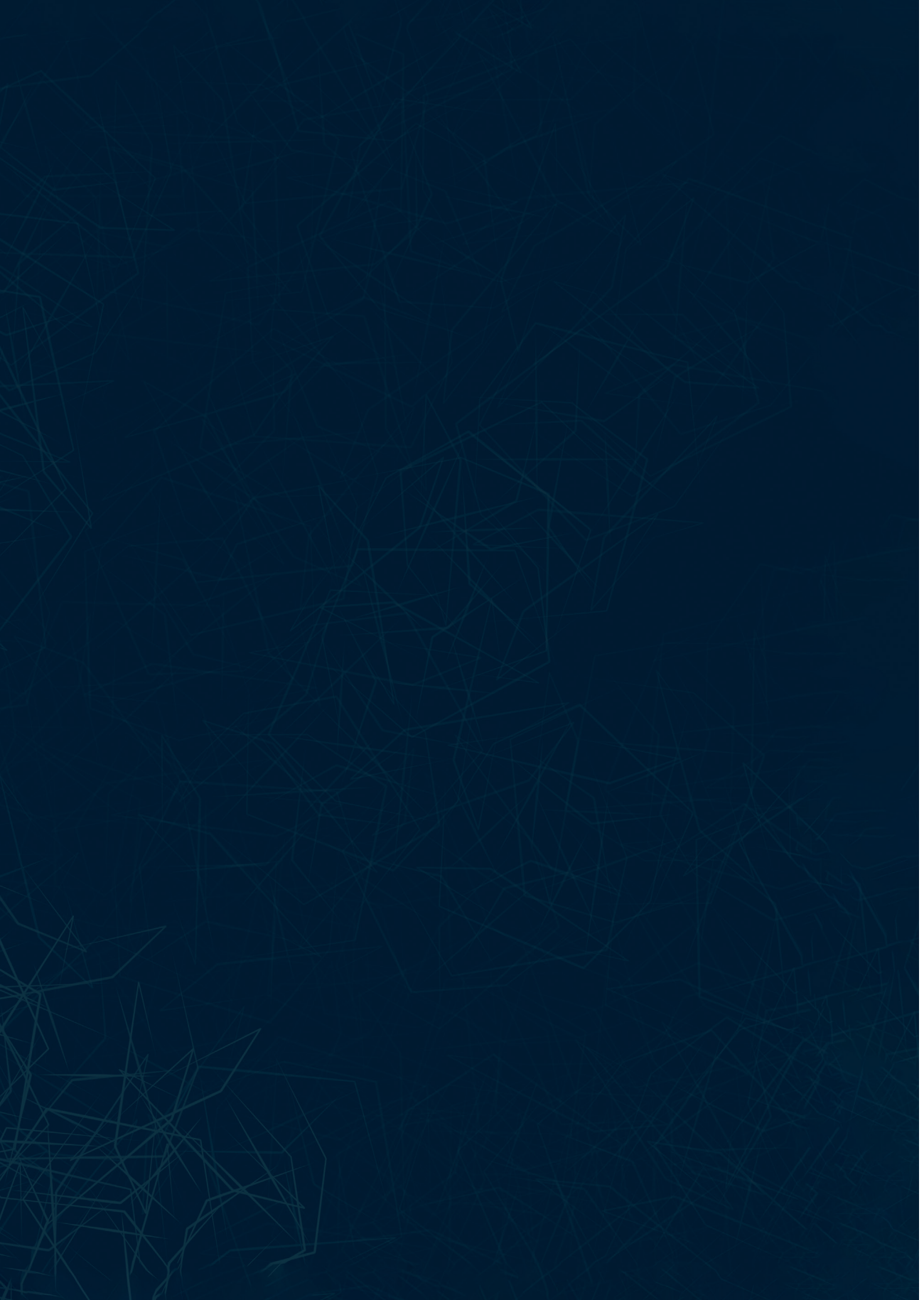
Prof. Dr. Philippe Coucke (University Hospital Liège)

Prof. Binsheng Zhao (Columbia University)


Dr. Alberto Traverso

Contents

Part I	Chapter 1	General introduction and outline of the thesis	9
	Chapter 2	Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework	21
Part II	Chapter 3	MRI-Based Radiomics Analysis for the Pretreatment Prediction of Pathologic Complete Tumor Response to Neoadjuvant Systemic Therapy in Breast Cancer Patients: A Multicenter Study	53
	Chapter 4	Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer	77
Part III	Chapter 5	Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods	103
	Chapter 6	The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset	149
	Chapter 7a	The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization	169
	Chapter 7b	Reply to Orhac and Buvat on "Ibrahim et. al, The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization Cancers 2021, 13, 13081848"	195
	Chapter 8	Reproducibility of CT-based Hepatocellular carcinoma radio-mic features across different contrast imaging phases: A proof of concept on SORAMIC trial data	205
	Chapter 9	MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability	225
	Chapter 10	Test-retest data for the assessment of breast MRI radiomic feature repeatability	245
	Chapter 11	MPenn radiomics reproducibility score: a novel quantitative measure for evaluating the reproducibility of CT-based handcrafted radiomic features	268
	Chapter 12	Deep learning-based classification of metastatic foci on bone scintigraphy	287
	Chapter 13	Validated fully automated detection and segmentation of non-small cell lung cancer on computed tomography images	303
Part V	Chapter 14	General discussion and future perspectives	331
Part VI	Appendices	Impact Paragraph	347
		Summary	349
		List of Publications	351
		Acknowledgements	354
		Curriculum Vitae	358



PART I

A large, textured blue splash with a white number 1 in the center. The splash is composed of various shades of blue, from light to dark, with a rough, organic edge. The number 1 is a simple, bold, white font, centered within the splash.

1

Chapter 1

General introduction
and outline of the thesis

Introduction

In recent decades, medical imaging has become a clinical cornerstone in diagnosing, managing and following up different diseases and clinical presentations [1,2]. Concurrently, much attention has been directed towards the applications of artificial intelligence (AI) on medical imaging with its different modalities, including computed tomography (CT), magnetic resonance imaging (MRI) and positron emission tomography (PET) scans. Advances in computational powers and machine learning algorithms, combined with the abundance of medical images, provided an opportunity for this field to grow exponentially [3]. These artificial intelligence methods include, but are not limited to, handcrafted radiomic features (HRFs) combined with machine learning, and deep learning (Figure 1) [4,5].

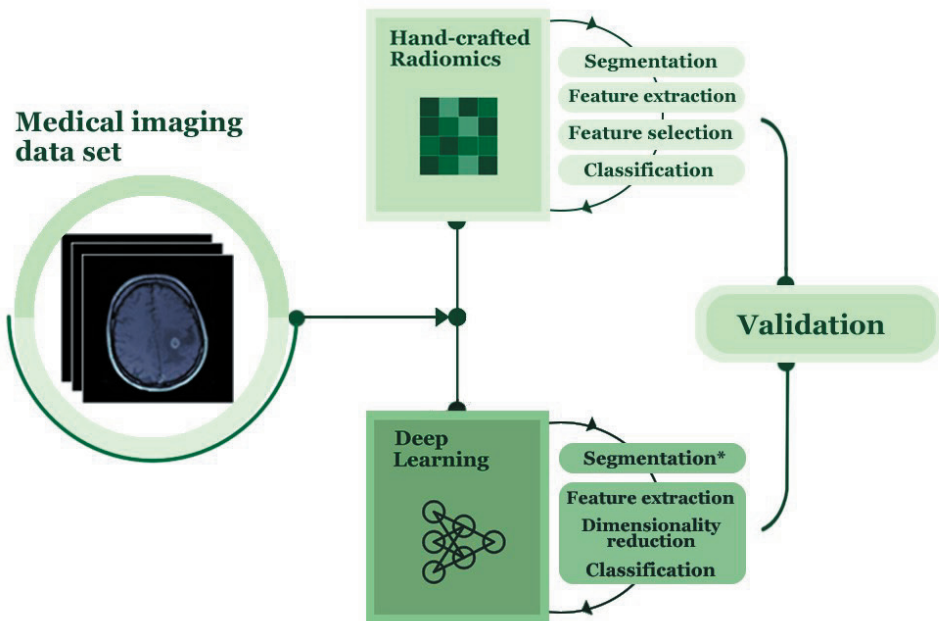


Figure I: Handcrafted radiomics and deep learning approaches (reprinted from Ibrahim et al.^[6]).

* Segmentation is not a necessity for DL approaches.

The main hypothesis motivating radiomics analysis is that quantitative imaging features decode biologic information of the region of interest (ROI) [6,7]. Both HRFs and DL provide possible alternatives to current clinical standards of care, since they can potentially provide fast, accurate, non-invasive and cost-effective means of clinical decision support, given that it has been extensively validated, and was developed while considering the sensitive nature of these techniques to variations in the medical imaging datasets.

Handcrafted radiomic features

Handcrafted radiomics refer to the high throughput extraction of quantitative features from medical images, which are then mined to look for correlations with biologic characteristics with the outcome being studied [8]. HRFs are extracted by applying predefined mathematical formulas on the array of values representing the medical image. An ROI is defined and HRFs are extracted thereof. A machine-learning algorithm is then used to develop a radiomic signature. Different groups of HRFs have been defined, including shape, intensity and texture features, and can be extracted from original and filtered images. These features are mined for correlations with biologic outcomes.

Many studies reported on the potential applications of HRFs-based signatures to predict patient outcomes, such as classification of lesions [9-11], response to therapy [12,13] and survival [8,14]. Nonetheless, a number of limitations has been identified [15,16]. A major identified limitation is the sensitivity of HRFs to temporal (test-retest) and image acquisition changes [17-19], which currently limits the translation of radiomic signatures to clinical practice. Figure 2 is an example of an HRF extracted from scans of the same volume of interest acquired with different imaging parameters. One scan was acquired with the standard kernel and a voxel size of $(0.39 \times 0.39 \times 1.25 \text{ mm}^3)$, while the other scan was acquired with the edge kernel and a voxel size of $(0.49 \times 0.49 \times 1.25 \text{ mm}^3)$. As illustrated, there was a large variation in the value of the HRF (NGTDM_Complexity), which are attributed to these variations in imaging parameters. The impact of these variations on the reproducibility of HRFs is not well understood. Moreover, several methods, such as image resampling and ComBat harmonization, have been used as potential methods to harmonize HRFs extracted from scans acquired differently. Nevertheless, the effects of these methods on the reproducibility of HRFs are yet to be investigated and understood.

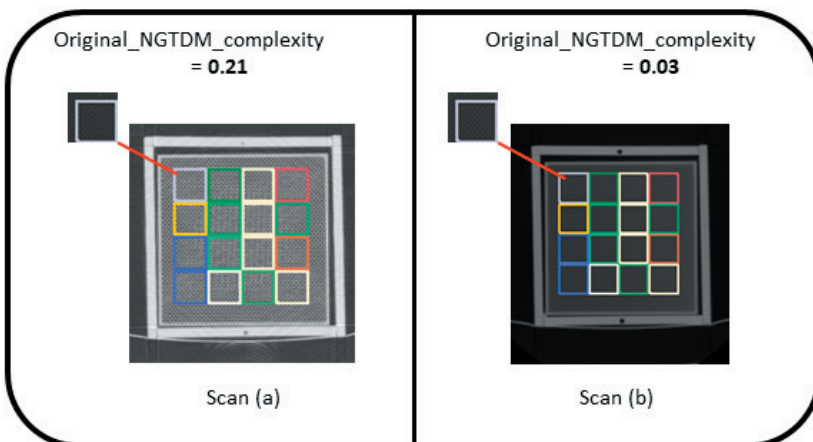


Figure 2: Thesis outline.

Deep learning

Deep learning is defined as data driven modeling techniques that employ the principles of simplified human neuronal interaction [5]. The artificial neuron models are the foundation units used to create complex chains of interactions named collectively as the DL layers. These layers are further combined to create the DL architectures. DL architectures are trained for recognition of problem related patterns in the data being analyzed to provide an automated tool to perform tasks. DL applications including, but are not limited to, automated segmentation [20,21] and classification [22,23] of medical images have become major focuses of AI research in medical imaging.

Deep features are more complicated in comparison to and some methods are being developed to help understand the mechanisms HRFs, and DL are generally considered a 'black box'. Yet, much attention has been paid to the explainability of DL models / factors based on which a DL algorithm makes a decision, such as Gradient weighted class activation mapping (Grad-CAM) method [24]. Grad-CAM generates an activation map that is superimposed on the original image to help visualize the reason for decision by DL algorithms.

Objectives, aims and Outline of the thesis

The overarching objective in this thesis was to gain more insights into the applications of HRFs and DL in medical imaging analysis, in an effort to guide developing AI-based tools as clinical decision support systems. For HRFs, based on the literature, the hypotheses that HRFs are subject to inter-reader variability, test retest variability, and are sensitive to variations in imaging parameters were tested. In addition, the hypothesis that a quantitative score can be used to assess the reproducibility of HRFs across scans acquired differently was tested. More specifically, the objectives were (i) to evaluate the current conventional HRFs workflow on scans acquired with different imaging parameters (Chapters 3 and 4); (ii) to investigate the effect of variations in imaging parameters on the reproducibility of features and some of the proposed methods to address the variations (Chapters 5-10); (iii) to develop a methodology to quantitatively assess the reproducibility of HRFs across scans acquired differently (Chapter 11). For DL, the main objective was to investigate the potential applications of DL to perform clinical tasks (Chapters 12 and 13).

The thesis is composed of 5 parts and 16 chapters. The description of these parts and chapters is below (see also figure 3).

Chapter I

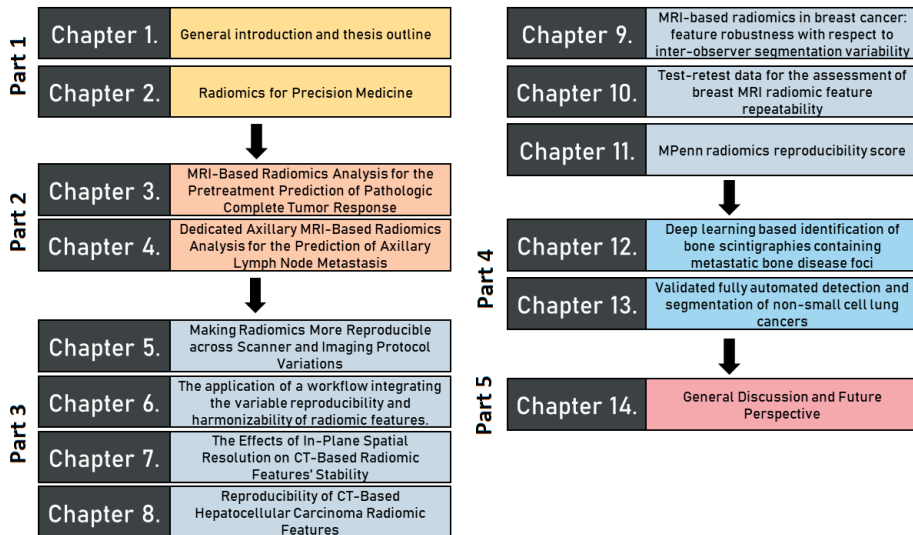


Figure 3: Thesis outline.

Part I: Introduction and a proposal of robust radiomic analysis framework

Chapter 1 is an introduction to the work carried out in this thesis, highlighting the main objectives of the work presented, and an outline of the thesis.

Chapter 2 serves as a general introduction to the applications of HRFs and DL methods in medical imaging. It is a review on the current applications of radiomics and the challenges radiomics currently faces. It further contains a proposal for a new radiomics framework that focuses on the reproducibility of HRFs based on the literature and previous experiments.

Part 2: Evaluation of the conventional handcrafted radiomics workflow

Chapter 3 is an experiment to assess the interpretability and generalizability of radiomic signatures developed using the conventional radiomics workflow on scans acquired differently. This work thoroughly investigated the modeling of a radiomic signature that can predict the complete pathologic response of breast tumors using HRFs extracted from breast MRI scans, in an effort to study the generalizability of radiomic signatures developed on scans acquired differently.

In **Chapter 4**, similar to Chapter 3, the work investigated the interpretability and

generalizability of radiomic signatures that can assess axillary lymph node status of breast cancer patients on MRI scans. The findings of this study further consolidated our conclusions derived in Chapter 3.

Part 3: The effects of variations in medical imaging on HRFs and validation of the proposed framework

Chapter 5 is a comprehensive review on the current harmonization methods used in radiomics analyses. It serves as a motivation for the investigation of potential of some of the harmonization methods to harmonize HRFs extracted from scans acquired differently.

Chapter 6 is an experiment that was aimed to investigate the reproducibility of CT based HRFs on phantom scans. The scans analyzed in this experiment (n=13) were acquired using different imaging vendors, models, acquisition and reconstruction parameters. The impact of ComBat harmonization on the reproducibility of HRFs was also assessed. It also serves as a validation for the framework proposed in Chapter 2.

In **Chapter 7**, an investigation into the effects of variations in in-plane resolution, while all other parameters are fixed, on the reproducibility of HRFs. Two sets of phantom scans, each composed of 7 phantom CT scans, were acquired similarly, except for the in-plane resolution. Concurrently, the impact of 10 different image resampling methods and ComBat harmonization on the reproducibility of HRFs was investigated. Additional analyses have been performed at the remark of another research group to further consolidate the findings and recommendations presented in the original study.

Chapter 8 describes a proof of concept on the reproducibility of HRFs extracted from CT based hepatocellular carcinoma HRFs in different imaging phases (arterial and portal venous phases). Furthermore, the potential of ComBat harmonization to remove the effects of differences in imaging phase has also been investigated.

In **Chapter 9**, the aim was to assess the inter-reader variability of MRI breast HRFs. In the experiment, a set of breast MRI scans were segmented by a number of medical doctors with varying experience in medical image segmentation. The agreement in HRF values extracted from these scans was then assessed to determine the sensitivity of HRFs to inter-reader variability.

Chapter 10 describes a test-retest experiment to assess the reproducibility of MRI breast HRFs. In this experiment, a number of healthy female volunteers underwent breast MRI scanning at two time points with multiple acquisitions at each time point. The study

further presents an analysis of the effect of different image preprocessing techniques on the reproducibility of HRFs.

In **Chapter 11**, further investigations into the collective effects of differences in different numbers of imaging parameters on the reproducibility of CT based HRFs are presented. A large set of phantom CT scans were analyzed. In addition, a novel score to assess the reproducibility of HRFs across CT scans acquired differently was developed based on the knowledge acquired from all the analyses performed.

Part 4: Some application of DL on medical images

Chapter 12 investigates the potential of a DL algorithm to detect metastatic bone disease on scintigraphy scans. This multicenter study included cancer and no-cancer patients, and the performance of the developed software was compared to that of uninformed nuclear medicine physicians in an in-silico trial. Furthermore, the explainability of the developed software was enhanced using Grad-CAM method.

Chapter 13 describes a DL software that can automatically detect and segment NSCLC on CT scans. The software consists of a pipeline that consists of several steps to ensure robustness of the software. The performance of the software was tested in an in-silico trial.

Part 5: General discussion and future perspectives

Chapter 14 serves as a general discussion regarding the work presented in this thesis, with recommendations and future prospects of the application of AI-based methods in clinical practice.

References

1. Zhang, X.; Smith, N.; Webb, A. Medical imaging. In *Biomedical Information Technology*; Elsevier: 2008; pp. 3-27.
2. Doi, K. Diagnostic imaging over the last 50 years: research and development in medical imaging science and technology. *Physics in Medicine & Biology* **2006**, *51*, R5.
3. Walsh, S.; de Jong, E.E.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T. Decision support systems in oncology. *JCO clinical cancer informatics* **2019**, *3*, 1-9.
4. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990)* **2012**, *48*, 441-446, doi:10.1016/j.ejca.2011.11.036.
5. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521*, 436.
6. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **2015**, *278*, 563-577.
7. Lambin, P.; Leijenaar, R.T.; Deist, T.M.; Peerlings, J.; de Jong, E.E.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.; Even, A.J.; Jochems, A. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **2017**, *14*, 749.
8. Aerts, H.J.; Velazquez, E.R.; Leijenaar, R.T.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications* **2014**, *5*, 4006.
9. Kumar, D.; Chung, A.G.; Shaifee, M.J.; Khalvati, F.; Haider, M.A.; Wong, A. Discovery radiomics for pathologically-proven computed tomography lung cancer prediction. In Proceedings of the International Conference Image Analysis and Recognition, 2017; pp. 54-62.
10. Li, H.; Zhu, Y.; Burnside, E.S.; Huang, E.; Drukker, K.; Hoadley, K.A.; Fan, C.; Conzen, S.D.; Zuley, M.; Net, J.M. Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TClA data set. *NPJ breast cancer* **2016**, *2*, 16012.
11. Choi, W.; Oh, J.H.; Riyahi, S.; Liu, C.J.; Jiang, F.; Chen, W.; White, C.; Rimner, A.; Mechalakos, J.G.; Deasy, J.O.; et al. Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer. *Med Phys* **2018**, *45*, 1537-1549, doi:10.1002/mp.12820.
12. Ha, S.; Park, S.; Bang, J.-I.; Kim, E.-K.; Lee, H.-Y. Metabolic Radiomics for Pretreatment 18 F-FDG PET/CT to Characterize Locally Advanced Breast Cancer: Histopathologic Characteristics, Response to Neoadjuvant Chemotherapy, and Prognosis. *Scientific reports* **2017**, *7*, 1556.
13. Bibault, J.-E.; Giraud, P.; Durdux, C.; Taieb, J.; Berger, A.; Coriat, R.; Chaussade, S.; Dousset, B.; Nordlinger, B.; Burgun, A. Deep Learning and Radiomics predict complete response after neoadjuvant chemoradiation for locally advanced rectal cancer. *Scientific reports* **2018**, *8*, 12611.
14. Oikonomou, A.; Khalvati, F.; Tyrrell, P.N.; Haider, M.A.; Tarique, U.; Jimenez-Juan, L.; Tjong, M.C.; Poon, I.; Eilaghi, A.; Ehrlich, L. Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy. *Scientific reports*

- 2018**, 8, 4003.
15. Yip, S.S.; Aerts, H.J. Applications and limitations of radiomics. *Physics in Medicine & Biology* **2016**, *61*, R150.
 16. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.; Dekker, A.; Fenstermacher, D. Radiomics: the process and the challenges. *Magnetic resonance imaging* **2012**, *30*, 1234-1248.
 17. Zhao, B.; Tan, Y.; Tsai, W.-Y.; Qi, J.; Xie, C.; Lu, L.; Schwartz, L.H. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific reports* **2016**, *6*, 1-7.
 18. Lu, L.; Ehmke, R.C.; Schwartz, L.H.; Zhao, B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PloS one* **2016**, *11*, e0166550.
 19. Xu, Y.; Lu, L.; Sun, S.H.; Lian, W.; Yang, H.; Schwartz, L.H.; Yang, Z.-h.; Zhao, B. Effect of CT image acquisition parameters on diagnostic performance of radiomics in predicting malignancy of pulmonary nodules of different sizes. *European Radiology* **2021**, 1-11.
 20. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the 3D Vision (3DV), 2016 Fourth International Conference on, 2016; pp. 565-571.
 21. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical image computing and computer-assisted intervention, 2015; pp. 234-241.
 22. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115.
 23. Wang, X.; Yang, W.; Weinreb, J.; Han, J.; Li, Q.; Kong, X.; Yan, Y.; Ke, Z.; Luo, B.; Liu, T. Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Scientific reports* **2017**, *7*, 1-8.
 24. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the Proceedings of the IEEE international conference on computer vision, 2017; pp. 618-626.

2

Chapter 2

Radiomics for precision medicine:
current challenges, future prospects,
and the proposal of a new framework

Authors

A. Ibrahim, S. Primakov, M. Beuque, H.C. Woodruff, I. Halilaj, G. Wu, T. Refaee,
R. Granzier, Y. Widaatalla, R. Hustinx, F.M. Mottaghy, P. Lambin

Adapted from

Methods. 2021 Apr 1;188:20-9

DOI

10.1016/j.ymeth.2020.05.022

Abstract

The advancement of artificial intelligence concurrent with the development of medical imaging techniques provided a unique opportunity to turn medical imaging from mostly qualitative, to further quantitative and mineable data that can be explored for the development of clinical decision support systems (cDSS). Radiomics, a method for the high throughput extraction of hand-crafted features from medical images, and deep learning -the data driven modeling techniques based on the principles of simplified brain neuron interactions, are the most researched quantitative imaging techniques. Many studies reported on the potential of such techniques in the context of cDSS. Such techniques could be highly appealing due to the reuse of existing data, automation of clinical workflows, minimal invasiveness, three-dimensional volumetric characterization, and the promise of high accuracy and reproducibility of results and cost-effectiveness. Nevertheless, there are several challenges that quantitative imaging techniques face, and need to be addressed before the translation to clinical use. These challenges include, but are not limited to, the explainability of the models, the reproducibility of the quantitative imaging features, and their sensitivity to variations in image acquisition and reconstruction parameters. In this narrative review, we report on the status of quantitative medical image analysis using radiomics and deep learning, the challenges the field is facing, propose a framework for robust radiomics analysis, and discuss future prospects.

Introduction

Advances in artificial intelligence applications, combined with those in medical imaging, have led to the gradual conversion of digital medical images into high-dimensional data appropriate for data mining and data science techniques ^[1]. Meanwhile, computing power and quantitative image analysis (QIA) techniques have made enormous progress, and the application of quantitative imaging techniques on medical imaging gained exponential momentum ^[2]. Currently, radiomics and deep learning are the most researched techniques on medical imaging.

Broadly, radiomics refers to the use of computational or statistical approaches to extract large numbers of quantitative features from a number of medical imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), to develop predictive models ultimately aiming to enable personalized clinical management ^[3-5]. Radiomic features are quantitative descriptions of the intensity, shape, volume, and texture of the region of interest (ROI), with the recent addition of more abstract features such as radial gradient and radial deviation ^[6]. Radiomics features are broadly divided into histogram-based and texture features. Different statistical methods are used to calculate the radiomics features. The methods include first-order statistics, which depends on the values of single voxels (histogram-based features for e.g. maximum and minimum intensity); second-order statistics, which depends on the relation between two voxels (for e.g. grey-level co-occurrence matrix (GLCM) features), and higher-order statistics (relations among three or more voxels, for e.g. neighborhood grey-tone difference matrices (NGTDM) features) ^[7,8]. The main hypothesis behind radiomics analysis is that radiomic features decode or correlate with the molecular characteristics, phenotype, and genotype of the region of interest (ROI) under study. This information can be used in combination with other patient information to improve patient management. Moreover, as the tumors are of heterogeneous nature ^[9,10], clinical approaches, such as tissue biopsies, might fail to characterize the entirety of the tumor ^[11]. In contrast, Radiomics takes the whole tumor region (or even the surrounding or healthy tissue) into account, which enables a better characterization ^[3]. Furthermore, frequent clinical imaging can transform radiomics into a non-invasive, easily repeatable, and cost-effective longitudinal approach for cDSS ^[12].

Deep learning (DL) is a field of data driven modelling techniques that utilizes the principles of simplified neuron interactions ^[13]. Using artificial neurons started to draw attention decades ago ^[14], but it only became a major research focus recently ^[15-17]. The artificial neuron model is used as a foundation unit to create complex chains of interactions - DL layers. These layers are used to generate even more complex structures - DL architectures. The neural network (NN) training procedure is typically a cost-

function minimization process. The cost function measures the error of predictions based on the ground truth labels ^[18]. Due to the high complexity of the network architectures, computational limitations are reached when trying to solve the optimization task analytically. Henceforth, iterative algorithms are used to overcome this issue. Commonly, these algorithms are variations of the gradient descent (GD). GD iteratively moves in the direction of steepest descent of the cost function, in order to find a local minimum. During the model training process, every image from the training dataset contributes to the cost minimization process. Thereby, a DL network learns how to solve a problem directly from existing data, and apply it to data it has never seen. These complex models contain the parameters (weights) for millions of neurons, which can be trained for the recognition of problem-related patterns in the data being analyzed. DL has been shown to be efficient in other fields, such as face recognition ^[19] and autonomous cars ^[20].

Since the introduction of the field, many studies have reported on the potential of such techniques for predicting patient outcomes ^[5,21,22]. The successful translation of QIA techniques into cDSS will have a significant impact on the clinical workflow and current patient management protocols. Clinicians will be able to non-invasively obtain a more detailed and accurate tumor characterization, in a shorter amount of time. Patients will have to go through less invasive procedures, while having treatment optimized based on their individual characteristics. Furthermore, patient-specific informed decisions can be made with more confidence. However, QIA is still developing in the field of medical imaging and several challenges, including the stability and reproducibility of imaging biomarkers, as well as the interpretability of the developed algorithms, need to be addressed before QIA can be translated to clinical applications.

In this narrative review, we focus on the current status of the potential of radiomics and deep learning to be incorporated in clinical decision support systems (cDSS), their challenges, as well as future prospects for these methods. We further propose a workflow to guide robust radiomics analysis.

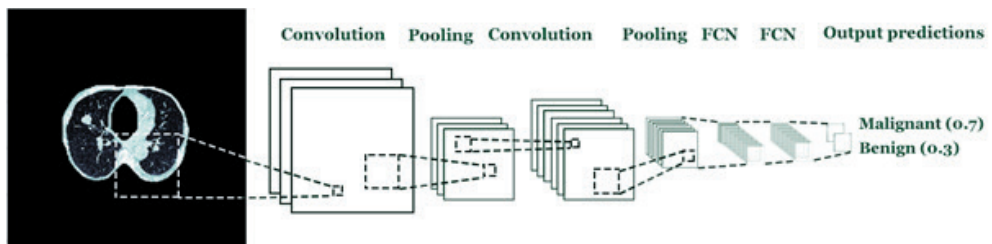


Figure I: Graphical depiction of DL architectures.

* FCN: fully connected network.

Quantitative image analysis for precision medicine

The need for personalizing the management of patients has been widely reported [23,24]. QIA represents a suitable candidate to be incorporated into the body of personalized medicine due to the non-invasive three-dimensional characterization of the ROIs, the availability of vast amounts of medical images, the longitudinal capabilities, and the cost-effectiveness of the method.

The currently implemented imaging biomarker development workflow is generalizable across different imaging modalities. The workflow can be described as consecutive steps divided into the main categories of data collection, image segmentation, features extraction, development of the signature, and evaluation of the performance (Figure 2), with the segmentation step being optional in the case of deep learning. The workflow has been previously extensively described [22,25].

Many studies have investigated and reported on the added clinical value of radiomics features for predicting various clinical outcomes, such as overall survival, tumor histology, response to therapy, and genetic profiling, among other endpoints. Furthermore, these studies were performed on various imaging modalities, including CT, MR, and PET.

While the hand-crafted radiomics pipeline necessitates the use of machine learning or statistical algorithms after feature extraction for modeling, DL techniques perform feature extraction and modelling internally without the need for further user interaction. DL has its own advantages and drawbacks compared to traditional radiomics. One of the key benefits of using DL is avoiding the contouring problem, the bottleneck of a traditional radiomics pipeline. However, due to the complexity of DL models, it is easier to overfit the model to the training data. As a result, a larger data set is needed for DL compared to hand-crafted radiomics. Furthermore, DL is considered a 'black box', i.e. the models and features generated are not (or barely) interpretable. This is currently one of the major challenges of the application of artificial intelligence (AI) in medical image analysis. Efforts are being made towards providing explainable AI algorithms, by investigating the correlation of the chosen features with biologic or semantic characteristics. Such correlations would provide an understanding about how the algorithm makes the decision, and ease its incorporation into cDSS.

QIA techniques have a great potential for involvement in developing classification, prognostic and predictive clinical tools. In comparison, classification tasks (for e.g. classifying tissue histology) seem to yield a better performance than predictive tasks (for e.g. survival prediction). This is in part due to the unaccounted for variables when trying to predict future events. In 2.1 and 2.2, we report on some examples that highlighted

the potential of radiomics and deep learning to predict various clinical endpoints, acknowledged or addressed the challenges of QIA techniques used, and/or applied the techniques on a relatively large sample size compared to other studies addressing the same clinical endpoint.

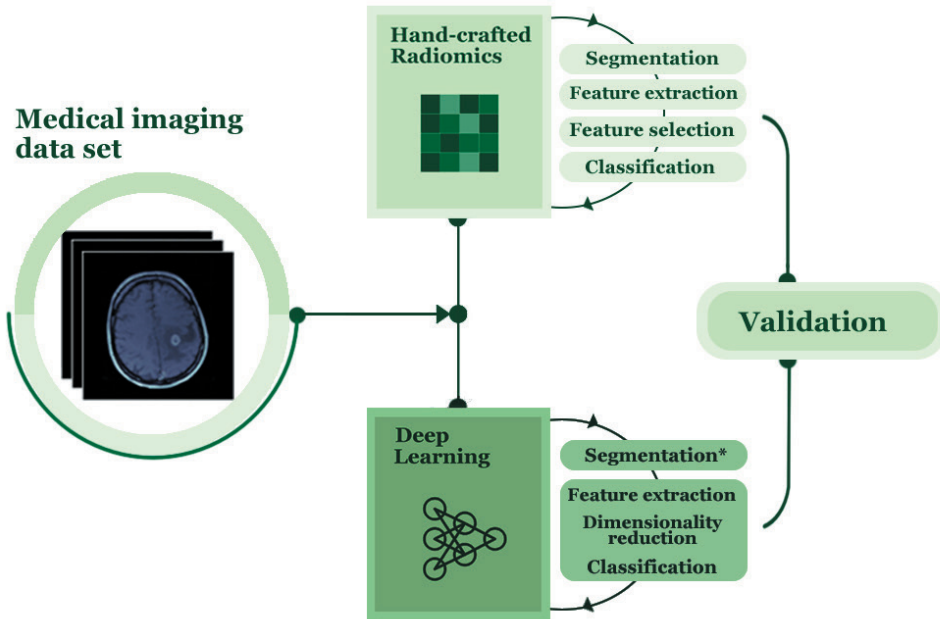


Figure 2: Development of imaging biomarkers using quantitative image analysis.

* Segmentation is not a necessity in the automated radiomics pipeline.

Hand-crafted radiomics

Overall survival

Wang et al. [26] investigated the potential of radiomics signatures to predict overall survival in patients with locally advanced rectal cancer. The authors tried to address the current clinical need for a risk stratification tool for such patients to safely forgo surgical resection, due to the high comorbidities associated. The study included 411 treatment planning CT-scans of patients treated with neoadjuvant chemotherapy followed by surgery. The authors developed a radiomics signature that could stratify patients into low- and high-risk survival groups. The radiomic features included in the signature were found to be independent of the clinical features. Adding radiomic features to the clinical model resulted in an improvement of the predictive power (c-index) of the clinical only model from 0.67 (0.62 - 0.73) to 0.73 (0.66 - 0.80) [26]. The authors used two investigations to ensure the selection of stable radiomics features, namely test-retest and contour-recontour robustness analysis. The results signifies the added value of properly using radiomics analysis on CT scans in improving patients' risk stratification. Yet, the

authors did not externally validate their signature, casting doubt on the generalizability of their signature. It is expected to be of value in cases where the scanning parameters are identical to those used in the study.

Another study by Bae et al. ^[27] investigated the potential of MR-based radiomics to improve the survival prediction of patients diagnosed with glioblastoma multiforme. The study is an effort to address the unmet clinical need for assessing the survival of the target group following therapy. The authors extracted radiomics features from 217 multiparametric MR scans of patients with glioblastoma. The authors identified 18 radiomics features to build a radiomic signature, and reported that the addition of radiomics features to clinical and genetic profiles of the patients significantly improves the stratification of patients ^[27]. The authors in this study applied a unique approach for the analysis by simultaneously analyzing radiomics features extracted from different co-registered MR sequences. The identified features were independent of the clinical and genetic factors, and the improvement in the survival prediction following their addition, supports the hypothesis of radiomics. Pitfalls in the study include the lack of assessment of radiomic feature stability before modeling, and as often seen in these studies, a lack of an external validation of the signature. However, their results support the hypothesis that radiomics are of great use when applied on scans acquired using identical settings.

Oikonomou et al. ^[28] reported on the potential of PET/CT-based radiomics to improve the survival stratification of patients with lung cancer treated with stereotactic body radiotherapy. The aim was to identify radiomic features that can improve the prognostication of patients following treatment. The authors extracted radiomics features from 150 PET/CT scans, and built radiomics signatures using 10 radiomics features. The authors reported that the radiomics signature was the sole predictor in the case of overall survival, and provided complementary information for the prediction of regional control ^[28]. The uniqueness in this study is the joint use of radiomics features extracted from the CT-component and PET-component of the PET/CT scans. The authors show how other currently used clinical parameters fail to predict overall survival, while only radiomics could. While the study highlights the potential of radiomics to improve risk stratification, no external validation of the signature was performed.

Progression free survival

Kirienko et al. ^[29] investigated the role of PET/CT-based radiomics to predict disease free survival in patients with non-small cell lung cancer undergoing surgery. The authors extracted radiomics features from PET, CT, and combined PET/CT images. The authors developed Cox regression models using only CT, only PET, and combined PET/CT radiomics features. They reported that the radiomic signatures they developed improve the current clinical stratification of the targeted patients ^[29]. The authors in this

study investigated the reproducibility of radiomics features across the different imaging parameters in their dataset. This ensured selecting the comparable features before proceeding with signature building. The authors also provide evidence of the added value of combining radiomics features extracted from different imaging modalities. Furthermore, the ability to predict disease free survival from the time of diagnosis -which radiomics offer- improves physicians and patients decision making. However, the authors in this study did also not perform an external validation of their signature. Further validation of the signature can prompt a prospective validation trial, before incorporation into cDSS.

Another study by Kickingereeder et al. ^[30] investigated the role of MR-based radiomics in predicting survival in patients with glioblastoma multiforme. The authors extracted radiomics features from 119 MR scans, and developed a radiomic signature using 11 features. The developed signature performed significantly better than the radiologic and clinical risk models, and its addition to those resulted in an overall improvement of progression-free survival stratification ^[30]. The finding that the radiomics signature performed better than the clinical and radiologic models supports the findings reported by Bae et al. ^[27], and adds more evidence that radiomic features decode complementary biologic information. However, the study did not address the issues of the reproducibility and generalizability sufficiently, leaving a room for improving the performance of radiomics.

Tumor histology

Wu et al ^[31] explored the role of radiomics in differentiating between the histologic subtypes of non-small cell lung cancer: adenocarcinoma and squamous cell carcinoma. The study was an effort to address the clinical need for less invasive and easily repeatable methods to determine tumor histology. The authors extracted radiomic features from 350 CT scans of NSCLC patients for whom the tumor histology has been determined from surgical specimens. The developed signature included 5 radiomics features, and they reported an area under the receiver characteristics curve (AUC) of 0.72 ^[31]. This study reflected on the potential of non-invasive radiomic signatures to differentiate between adenocarcinoma and squamous cell carcinoma. They also investigated different machine learning methodologies for building the radiomics signature. While this study generates evidence for the potential of radiomics, the performance of the developed signature is significantly lower than the current gold standard -tissue biopsy. However, there is a great room for improving the development and performance of the signature. The authors did not address the acknowledged challenges in radiomics, nor did they validate their signature on an external dataset. Preselection of reproducible features, external and prospective validation of the signature are necessary steps in the development of radiomics biomarkers.

In another study, Wu et al. [32] investigated the added value of MR-based radiomic features for the prediction of hepatocellular carcinoma (HCC) grade. The authors extracted radiomic features from 170 MRI scans of HCC patients, whose tumor grade was identified through pathological samples. The radiomics-only signature (AUC of 0.74) outperformed the clinical model (AUC of 0.60), and the combination of both significantly improved the prediction (AUC of 0.80) [32]. The authors in this study also combined radiomic features extracted from two different MR sequences and analyzed them simultaneously. The significant improvement of the predictions following the combination of clinical and radiomic features supports the independence of radiomics features from other clinical information. However, external validation of the developed signature is still a necessity before confidently performing prospective validation.

Valleries et al. [33] explored the potential of the combination of FDG-PET- and MR-based radiomics features to classify lung nodules. The authors extracted radiomics features from 51 PET and MR scans of histologically confirmed lung lesions in patients with soft-tissue sarcoma. The authors achieved a sensitivity of 0.96 and specificity of 0.93 in diagnosing metastatic nodules using a model with combined radiomic features from both PET and MR modalities. The authors used a novel interesting approach by simultaneously analyzing the features extracted from FDG-PET and MR scans, and were the first to show the potential of this method. The performance of the developed signature makes it a suitable alternative for patients for whom tissue biopsy is contraindicated. Its possible translation to cDSS might significantly improve patient outcomes, as treatment is based on the histologic diagnosis. Yet, further external and prospective validation of the signature is needed.

Response to therapy

Trebeschi et al. [34] explored the role of radiomics in predicting response to anti-PD1 immunotherapy in patients diagnosed with advanced melanoma and NSCLC patients. Immunotherapy has shown promising results. Yet, there is still a need for a tool to determine which patients will benefit from receiving anti-PD-1 antibodies. The authors extracted radiomic features from 1055 ROIs segmented on 203 CT scans. The authors developed a radiomic signature that could predict the response to therapy with an AUC of 0.76; showing the potential of radiomics to predict response to therapy in such patients [34]. Interestingly, the authors found correlations between the radiomic biomarker and the genes associated with cell cycle progression and mitosis. Radiomics can become a tool for assisting decision making in immunotherapy, a great unmet clinical need. The study however did not externally validate the signature, and did not sufficiently address the issues of feature stability and reproducibility. Therefore, the application of the developed signature is also limited to the patients who are scanned with the same scanning parameters as used in the training.

In a study by Horvat et al. ^[35], the authors investigated the role of radiomics in assessing complete clinical response (cCR) after neoadjuvant chemoradiotherapy (CRT) in patients with locally advanced rectal cancer. The guidelines of treating these patients include surgery, but evidence showed recently that a select group of patients can be safely treated with only CRT. The authors extracted radiomic features from 114 MR scans, and developed a radiomics signature with a sensitivity of 1.00, and a specificity of 0.91, which outperformed qualitative assessment of the response performed by two radiologists. The current clinical standard evaluation of cCR includes digital rectal examination and endoscopy, with an accuracy ranging between 0.71 and 0.88 ^[35]. The developed radiomic signature showed the highest accuracy among the available compared-with tools. Nonetheless, several steps to improve the methodology and performance of the radiomics signature could be made. The sound cCR evaluation following RCT can improve the patient management by eliminating surgical risks, time and money.

Deep learning

The application of deep learning on medical imaging could potentially fulfil more complicated tasks than hand-crafted radiomics, especially when large amounts of data are available. Furthermore, as definition of the ROIs is not a necessity in the automated deep learning workflows, the algorithm will learn patterns from the whole image and possibly make connections with the habitat of the ROIs. The applications of neural networks on medical imaging are also not limited to classification and prediction of clinical end points, but can extend to include other tasks, such as the detection and segmentation of abnormalities or target volumes, which have been investigated for decades ^[36]. Especially the detection and segmentation of lesions can be easily incorporated into the radiomics workflow, further automating the process. In the following paragraphs, we give examples of different applications of DL on medical imaging to perform various tasks on datasets acquired with one of the three main medical images modalities: CT, MRI, and PET.

Automatic segmentation of target structures

Jiang et al. ^[37] tried to develop a DL model that is able to accurately perform volumetric lung tumor segmentation on CT images. The authors used two versions of multiple resolution residual network models for the delineation of the ROIs. The authors used 377 tumors from the open source dataset available on The Cancer Imaging Archive (TCIA) (<https://www.cancerimagingarchive.net>) to train the model, and two independent datasets of 304 and 529 lung tumors to validate it. The dice similarity coefficient (DSC), which measures the spatial overlap of the segmentations, was computed to evaluate the performance of the model. The DSCs of the model on the two validation datasets were 0.75 and 0.68, respectively. The authors reported that there was no significant difference between the DL-generated mask and experts' segmentations ^[37]. The new approach for

segmenting medical images used in this study shows to be superior to the traditional use of UNet. The approach generalizes well on external data and overcomes the multiple sizes problem. The major pitfalls is that the authors did not use the 3D geometry of the CTs to compute the results, which would probably increase the performance significantly. The translation of such a tool to clinical practice will significantly reduce the time spent by the clinicians to plan the treatment, or evaluate the response to therapy. Moreover, from a research perspective, it can significantly reduce the time needed for radiomics research, and it will address the issue of inter-observer sensitivity of radiomics features.

In the study by Yi et al. [38], the authors developed a DL model for the segmentation of brain tumors based on 274 brain MRIs extracted from the Brain Tumor Image Segmentation Benchmark (BRATS) dataset [39]. Segmentation of brain Glioblastoma on MRI is a time-exhaustive process, and an automated, accurate and reproducible tool for this purpose is considered a clinical need. The model was trained using four different MRIs sequences. The particularity of their convolutional neural network (CNN) model is a fixed difference of Gaussian filters as a first convolution layer, as it was proven to be the most efficient for 3D segmentation. The DSC for the model was 0.89 on the BRATS dataset when compared to ground truth segmentations [38]. This article shows the superiority of 3D CNN compared to 2D CNN. The algorithm generated segmentations with a volumetric overlap of 0.89 with the experts' segmentations, which shows the potential of these tools for clinical use. However, the lack of external validation in the study limits the applicability of the algorithm to scanning parameters in the training set. The clinical practice can benefit from such tools, as it significantly reduces the time the clinicians spend, and can provide more accurate evaluation of tumor response than the current clinical routine.

Chen et al. [40] explored the possibility of developing a DL model that is able to detect and segment cervical tumors on PET imaging. The authors proposed prior information constraint CNN (PIC-CNN), which integrates a CNN with prior information of cervical tumor. The authors reported a DSC of 0.84, which was superior to the other methods in the comparison, including transfer learning based on fully convolutional neural networks (FCN) (DSC of 0.77), automatic thresholding (DSC of 0.59), and region growing method (DSC of 0.52) [40]. The study highlights the potential of deep learning to perform well-defined and robust segmentations on PET imaging. The novelty of the approach is the use of prior information as input of the model, with delineation of the bladder. This extra information seems to give the traditional model an advantage compared to models that solely segment the tumors. However, the results were not validated on an external dataset. The application of the developed algorithm -after validating it- would decrease the need for tissue biopsy, as well as the time spent on segmenting the tumors manually or semi-automatically.

Oncologic Classification tasks

Ardila et al. ^[41] tried to predict the risk of lung cancer using screening low-dose CTs. The algorithm is trained on screening low-dose CT scans of patients who were known to be at risk. The authors trained their DL model on approximately 7000 scans, and validated its performance on 1139 cases. The authors reported that the model achieved the “state-of-the-art” performance (AUC of 0.944). Furthermore, the model outperformed all the radiologists (n=6) who were asked to give predictions. The model resulted in a significant reduction in the false positive (11%), and false negative rates (5%) ^[41]. While the current low-dose CT screening protocol has substantially improved in terms of consistency, it still faces major limitations represented in the inter-observer variability and incomplete characterization of image findings. The authors in ^[41] developed an algorithm that achieved significantly better performance than the current protocol, highlighting the potential of DL algorithms to revolutionize the field of lung cancer screening. Other advantages of the algorithm are that it eliminates the current clinical practice limitations.

Ismael et al. ^[42] investigated the ability of DL algorithms to classify different brain tumors. The algorithm predicts if the lesion is one of: Meningiomas, Gliomas, and Pituitary tumors. The authors developed the algorithm on 3064 T1 MRI images from 233 cancer patients. As input to the algorithm, the 2D images were considered independent from each other, and were split into 80% training and 20% testing, with strictly different patient data. The classifier used is ResNet50, a classic deep learning network, and the resultant balanced accuracy was 0.99 on a slice level and 0.97 at a patient level. This study shows that deep learning can very accurately classify brain tumors based solely on MRI data. However, the data to be used should be acquired using the same scanning parameters, as no external validation was performed in this study. There is a great clinical significance from the development of such a cDSS, as it eliminates the need for risky brain biopsies, while maintaining high accuracy.

In another study by Sibille et al. ^[43], the authors used the combination of CT, fluorine 18-fluorodeoxyglucose PET, atlas and PET maximum intensity projection (MIP) imaging to classify lung nodules. The study included a set of 629 patients who were diagnosed with either lung cancer or lymphoma. The authors developed models using each of imaging modalities separately, as well as in combination. The recommended algorithm achieved an AUC of 0.98 when both CT and PET were combined ^[43]. This study shows that the combination of CT and PET can achieve an outstanding performance in terms of predictions. The current clinical practice requires the clinician to review and classify all of the increased-uptake foci in a PET/CT scan. The algorithm could help the clinicians to quickly read those images, after highlighting the suspicious areas and their most likely classification using DL.

Non-oncologic Classification tasks

Walsh et al. ^[44] explored the potential of DL to classify fibrotic lung diseases using high resolution CT scans. The current clinical guidelines for classifying fibrotic lung diseases are based on high resolution scans, and diagnoses are made based on the semantic features identified by the radiologists. While these guidelines are the current gold-standard, it suffers greatly from inter-observer variability. The authors tried to address this unmet clinical need using DL approaches. The authors trained their DL model on 929 CT scans, and tested it on 139 scans. The authors reported a performance with human-level accuracy (0.76) ^[44]. Of interest, the algorithm developed had a better agreement with expert radiologists than among them. The ease of application of such methods in clinical settings could benefit clinical practice, especially in centers where such clinical expertise is scarce.

In the study by Ding et al. ^[45], the authors tried to develop a DL model that is able to diagnose Alzheimer's disease (AD), using ¹⁸F-FDG PET scans of the brain. The current clinical guidelines to diagnose AD necessitate the interpretation of scans by an expert, and there is no definitive biomarker. To investigate the potential of DL, the authors collected two datasets: one used for training and testing the model (n=2109 scans), which was split into 90% training and 10% testing; and an independent dataset (n=40) for the validation of the model. The authors reported an AUC of 0.98, sensitivity of 1.00 and specificity of 0.82, using scans acquired 75.8 months on average before establishing the diagnosis. The model further outperformed the readers' performance (sensitivity of 0.57 and specificity of 0.91) ^[45]. The significance in this study lies within the novelty of developing a biomarker for AD that is currently an unmet clinical need. In addition to the significantly better performance compared to human experts, the model can predict that the patient has AD in progression significantly earlier (-6 years). Such an application will revolutionize the clinical management of AD. However, prospective validation of this signature is needed before its translation to clinical practice.

Oh et al. ^[46] applied a DL based approach in order to classify the neuroimaging data related to AD. Authors used 694 MRI scans (T1-weighted MP-RAGE sequence) for solving several binary classification problems: AD vs. normal control (NC), progressive mild cognitive impairment (pMCI) vs. NC, stable mild cognitive impairment (sMCI) vs. NC and pMCI vs. sMCI. The authors utilized convolutional autoencoder-based unsupervised learning algorithms in order to classify the AD vs. NC. Following that, the authors applied a supervised transfer learning approach to classify the pMCI vs. sMCI. The developed algorithms achieved accuracies of 0.87, 0.77, 0.63, and 0.73 for the AD, pMCI, sMCI and pMCI vs. sMCI classifications, respectively. In comparison to Ding et al. ^[45], the authors in this study used different DL approaches, and less numbers of patients were available for training and testing the algorithm. Furthermore, the

difference in the imaging modality analysed in each study could justify the variation in performance, as AD begins with functional impairment rather than structural changes. Although the model developed by Oh et al. ^[46] was outperformed by human experts, the authors demonstrated the possibility of end-to-end DL algorithms, which could be translated to clinical use after further optimization and prospective validation.

Response to therapy

Lou et al. ^[47] reported on the potential of DL models to predict response to radiotherapy in patients with lung cancer (primary or metastatic) using CT scans. Currently, all patients are treated similarly, while personalizing radiotherapy remains a desired, but unmet clinical need. The authors in this study collected a total of 849 scans for training the DL algorithm, and 95 scans to validate it. The authors developed a deep learning model (deep profiler) that computes and includes radiomic features in the deep-profiling process. A model combining the deep profiler and clinical variables is then used to calculate a risk score that is used to predict the response to treatment. The algorithm classifies patients into high and low risk groups, with a high performance (c-index of 0.72), which is significantly better compared to the results obtained with solely handcrafted radiomic models (c-index between 0.65 and 0.68) ^[47]. The algorithm developed in this study opens new potentials for individualizing radiotherapy based on patients' sensitivity. Thereby, avoiding over- or under-treatment, and the side-effects of unnecessary treatment. Nevertheless, proper prospective validation of the developed algorithm remains a necessity.

Ypsilantis et al. ^[48] used convolutional neural networks to develop a model that is capable of predicting response to neo-adjuvant chemotherapy (NAC) in patients with esophageal cancer using PET scans. NAC is considered a standard of care in some cancers. While NAC has favourable outcomes in patients who respond, patients who do not end up with worse outcomes. To investigate the potential of QIA techniques, the authors collected 107 PET scans of patients diagnosed with esophageal cancer, treated with NAC, and followed-up to determine response. The authors compared the performance of hand-crafted radiomics with deep learning approaches. The authors reported that the developed deep learning algorithm outperformed the hand-crafted radiomics model, and achieved a sensitivity of 0.81 and specificity of 0.82 ^[48]. The algorithm developed in this study highlights the potential of using DL to predict patients' response to therapy at baseline, which is considered a substantial clinical added value.

Challenges and future directions

Biomarkers are defined as “objective indications of medical state observed from outside the patient – which can be measured accurately and reproducibly” ^[49]. The core

of choosing a biomarker is the ability to measure it objectively. The reproducibility of imaging quantitative features across different imaging parameters is currently the steepest hurdle in QIA. As more research is being performed, other challenges, such as the sensitivity of QIA features to variations in the segmentation of the ROIs; and the lack of feature reproducibility across different implementations of radiomics toolboxes, are becoming increasingly clear.

The stability and reproducibility of quantitative features

Since the first landmark study in radiomics by Aerts et. al [50], the sensitivity of radiomic features to repeated acquisitions has been acknowledged. The authors performed a test-retest stability investigation, and used 100 out of 440 calculated radiomic features based on the stability rank of the features. The authors also acknowledged the sensitivity of features to differences in segmentations, and performed a primary feature selection based on the features' robustness with regards to differences in both test-retest and segmentations. More recently, several studies reported on the sensitivity of radiomic features to temporal changes in test-retest studies across different modalities, including CT, MRI, and PET.

Anatomical imaging

Anatomical imaging (CT and MRI) is used to explore the underlying anatomical structures. CT imaging is standardized using the hounsfield units (HU) [51]. On the other hand, MR imaging has no such standardized intensity measurements [52]. Even though CT imaging uses standardized measurements, CT-based radiomics are not necessarily reproducible. Several studies reported that a significant number of CT-based radiomic features are not reproducible in test-retest settings, where the scans are acquired using the same scanning parameters [53–55]. Other studies that investigated the reproducibility of CT-based radiomics features across different imaging acquisition and reconstruction parameters reported that the majority of radiomic features are significantly affected by such differences [53,56,57]. Unreproducible radiomic features should be removed before starting the modeling of radiomics signatures. Therefore, it is always necessary to perform preselection of stable radiomics features based on the data under study, before starting the modeling.

MR-based radiomics is even more complex and challenging to standardize compared to CT based radiomics, as more factors -in addition to lack of standardized intensity measurements-affect MR imaging [58]. Some studies reported on the stability of various MR-based features. Fiset et al. [59] investigated the reproducibility of T2-weighted MRI of cervical cancer in three different settings: (i) test–retest; (ii) diagnostic MRI versus simulation MRI; (iii) interobserver variability. The authors reported that 22.6%, 6.2% and 74.4% of 1761 extracted radiomics features were reproducible across test-retest,

diagnostic versus simulation MRI, and different observers, respectively. Semi-parametric maps derived from specialized MRI sequences suffer less from the lack of stability: Peerlings et al. ^[60] reported on the stability of radiomics features extracted from apparent diffusion coefficient (ADC) map in test-retest and across different cancer types, centers, and vendors. The authors reported that out of 1322 extracted radiomics features, 122 features were stable across all cancers, centers, and vendors.

On top of these challenges, using contrast agents for imaging adds another level of complexity to the reproducibility of features, due to the differences in the cardiac function of patients being scanned. Changes in cardiac function can affect the time the distribution of the contrast in the body takes ^[61]. Another factor in contrast-enhanced images is the difference in time between the injection of the contrast and scan acquisition, which might be slightly different across centers and protocols.

Functional imaging

Functional imaging is used to assess the metabolic activity of a region of interest, and includes the injection of radiopharmaceuticals. Some standardized measurements in PET are already being extracted and used in clinical practice, such as the standardized uptake value (SUV), and the metabolically active tumor volume (MTV) ^[7].

The challenges of radiomics for functional imaging are similar to the challenges of contrast-enhanced anatomical imaging radiomics, where the variability in the injected radiopharmaceutical activity, the time between injection and image acquisition, and acquisition time per bed position have profound implications on the reproducibility of radiomics features ^[62]. In addition, functional imaging lacks anatomical specificity and suffers from low resolution, which could be addressed by the use of hybrid imaging ^[22]. Tixier et al. ^[63] investigated the reproducibility of SUV measurements, intensity histogram features, intensity-size zone features, and co-occurrence matrices features. The authors acquired two ¹⁸F-FDG PET scans of 16 patients, with a 4-days' time interval. In contrast to further studies, the authors reported that these features were insensitive to the discretization range. Hatt et al. ^[64] investigated the robustness of PET based heterogeneity textural features with respect to the delineation of functional volumes and partial volume effects correction. The authors reported that these features were significantly affected by the differences in the delineation. The authors further reported that local features, e.g entropy and heterogeneity, were more robust when compared to regional features, e.g intensity variability and size-zone variability. Leijenaar et al. ^[65] investigated the role of SUV discretization on radiomics features. The authors used two different methods for SUV discretization, and reported that differences in SUV discretization significantly affect the reproducibility of ¹⁸F-FDG PET based radiomic features. The authors recommended the standardization of methodology for radiomics

analysis. Altazi et al. ^[66] investigated the reproducibility of PET based radiomic features in cervical cancer patients. The authors investigated the reproducibility in three different scenarios: (i) manual versus computer-aided segmentations, (ii) gray-level discretization, and (iii) reconstruction algorithms. The authors extracted 79 PET radiomics features, and reported that the percentage of stable features in the three scenarios were 13%, 5%, and 1% respectively. Shiri et al. ^[67] explored the effects of different reconstruction on ¹⁸F-FDG PET radiomics. The authors studied the effects of several factors including number of sub-iterations, number of subsets, full width at half maximum (FWHM) of Gaussian filter, and scan time per bed position and matrix size. The authors reported that 47% of the features were found to be robust, and these include shape, 44% of the intensity based features, and 41% of the texture based features. However, with changes in matrix size, the authors reported that only 6% of the features were robust.

The discrepancies in the reported percentages of stable/reproducible features across the reported studies are most likely linked to the variations between the datasets used in each of the studies in the scanners, and scans acquisition and reconstruction parameters combinations. However, these discrepancies are expected because of the different complexity of radiomics features, as well as the interaction between the different scanning parameters. All of the above mentioned studies reported that a variable percentage of radiomics features are affected, which highlights the necessity of performing feature stability/reproducibility studies based on the data under analysis before performing radiomics analysis.

Sensitivity of quantitative imaging features to variations in the segmentation of the ROIs

In QIA, the medical images are converted to numerical arrays before feature calculation. Consequently, it is intuitive that differences in segmentations affect the quantitative imaging feature values variably, depending on the feature definition. Many studies have identified lists of radiomics features that are robust to variability in segmentations ^[50,68,69]. Furthermore, with the inclusion of deep learning methods in image analysis, efforts are being made to develop reliable and reproducible automatic segmentations of various regions of interest as described in 3.2.1. Deep learning suffers less in this aspect, as the provision of ROIs is not obligatory.

The different implementations of radiomics feature extraction toolboxes

It is common knowledge in the radiomics community that different radiomics toolboxes use different pre-processing techniques and/or feature definitions, which lead(s) to variations in estimation of radiomics feature values when different software solutions are used. To address this issue, the radiomics community started an initiative – Imaging Biomarkers Standardization Initiative (IBSI) - that aims at standardizing radiomics feature extraction using different toolboxes ^[70]. To date, the IBSI standardized the

extraction of 169 radiomics features ^[71]. Limiting the radiomics analysis to the IBSI standardized features can facilitate radiomic features interchangeability across platforms.

Future directions

To address the issue of radiomic features reproducibility, some harmonization methods have been investigated in the literature. Of the trending methods is Combine Batches (ComBat). ComBat is a statistical method that was developed to remove the batch effects in microarray expressions ^[72]. While several studies have reported on the application of ComBat harmonization in radiomics analysis as a means to remove batch effects ^[73,74], its direct application on radiomics data is not in concordance with the mathematical definition of ComBat ^[72], or with the hypothesis that radiomics correlate with biology. This is because ComBat assumes that the differences between batches are attributed to two groups of factors, the first group refers to the biological covariates, which radiomics features are investigated for correlations with. Moreover, adding biologic covariates for ComBat in the training of radiomics signatures will hinder its prospective use, because it will be the outcome the radiomic signature tries to predict. The second group refers to the “non-biologic” factors, such as image acquisition and reconstruction parameters. ComBat was defined to handle one batch effect at a time. In contrast to gene expression arrays for which ComBat was designed, radiomic features have different complexity levels, which are expected to be non-uniformly affected by the variations in imaging parameters. In addition, the differences in image acquisition and reconstruction settings in a given retrospective imaging dataset are usually in more than one imaging parameter. The proper use of ComBat would require the assessment of the reproducibility of radiomics features after applying ComBat on representative objects with no biologic variations, such as phantoms. Then, radiomics features extracted from patients’ scans acquired with the same imaging parameters can be transformed based on the location/scale parameters estimated by the application of ComBat on the phantom data. We here propose a framework for performing robust radiomics analysis (Figure 3). Nonetheless, a radiomics-specific harmonization method is still needed to eliminate the need for phantom studies, as the performance of ComBat is expected to be dependent on the variations in scanning parameters in the data.

The workflow consists of consecutive steps, and can be used to preselect reproducible and harmonizable radiomics features. The first step in the workflow is the collection of retrospective patient imaging data to be analyzed. In the second step, scan acquisition and reconstruction parameters must be extracted from the collected patient data. The next step includes scanning a phantom with the parameters extracted from the patient imaging data. This allows the assessment of the reproducibility of radiomics features across the different scan acquisition and reconstruction parameters, and the selection of those features for performing robust radiomics analysis.

Based on our review of existing literature and our own experience, in order to use ComBat in the context of radiomics analysis (steps 5-7), two extra steps are needed. After selecting the features that are insensitive to the variations in the scanning parameters extracted from the patient data, features that are reproducible in test-retest in each of the combinations of those scanning parameters must be identified. ComBat is then applied on the features that are reproducible in test-retest but not across different scanning parameters. The concordance of Radiomic features is assessed following the application of ComBat. The location/scale shift parameters estimated by performing ComBat on the phantom data are then applied to the radiomics features extracted from patient data to harmonize them. The combination of the identified stable and harmonizable features can be further used to build the radiomics signature.

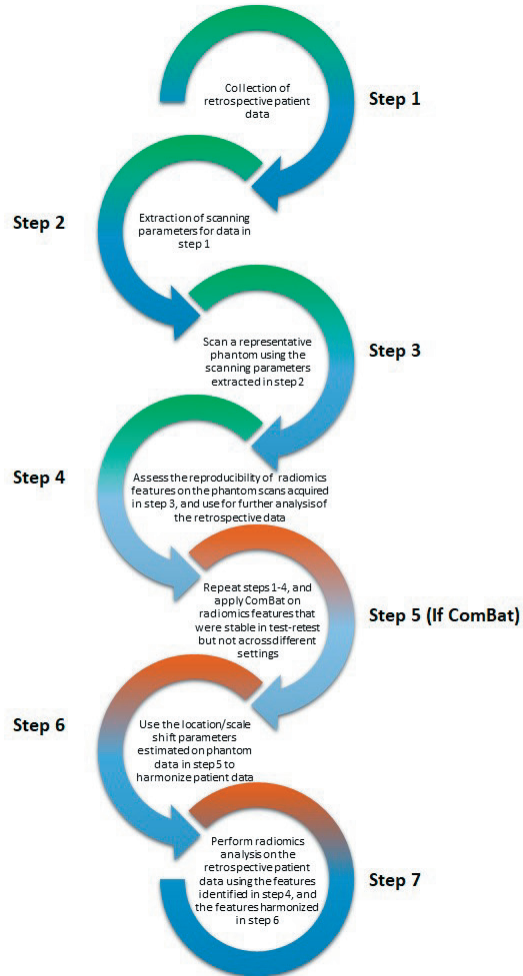


Figure 3: Proposed workflow for robust radiomics analysis.

The challenges discussed above raise questions about the future applications of radiomics, and the development of radiomic signatures as clinical biomarkers. To begin with, how to approach the concept of external validation in radiomics studies. Do radiomic signatures need to be externally validated as is the case with other biomarkers, given all the challenges of reproducibility across different imaging settings? Or would the observational prospective validation of a given signature in a specific image setting suffice? Does the development of radiomic signatures need to be specific for a scanner model and imaging settings? The ultimate solution will be the development of specific quantitative imaging parameters, as there is currently a clinical direction to personalize imaging settings per patient, which will have its toll on radiomics analysis.

Conclusion

Quantitative imaging techniques (radiomics and deep learning) present a perfect candidate for personalizing patients' management. Applying these techniques in a sound manner can provide highly accurate and reproducible tools that minimize costs and time loss. However, to incorporate QIA in cDSS, the quantitative features should fulfil the definition of a biomarker, namely the stability and reproducibility. The future of quantitative image analysis in general lies within harmonizing the imaging protocols across centers and scanners, or within the development of a unique global protocol for quantitative analysis scans. Hence, the development of radiomics-specific tools to harmonize medical images and facilitate meaningful quantitative image analysis of the currently available retrospective data remains a necessity. Our proposed framework is expected to improve the robustness of radiomics analysis. Nevertheless, the benefits of the proper application and translation of QIA on medical imaging are undoubted. QIA techniques will be a valuable asset for both: the clinicians and patients. QIA can become an efficient means for aiding clinicians in risk stratification, early diagnosis, and improved management of patients.

Authors' contributions

A. Ibrahim: conceptualization, methodology, formal analysis, data curation, writing-original draft, project administration. **S. Primakov:** formal analysis, data curation, writing- original draft, visualization. **M. Beuque:** formal analysis, data curation, writing- original draft. **H. Woodruff:** Supervision, writing-review and editing. **G. Wu:** Resources, data curation. **T. Refaee:** Resources. **R. Granzier:** Resources. **I. Halilaj:** visualization. **Y. Widaatalla:** Resources. **R. Hustinx:** Supervision. **F.M. Mottaghy:** supervision, writing-review and editing. **P. Lambin:** Conceptualization, methodology, writing-review and editing, project administration, supervision.

References

1. S. Walsh, E.E.C. de Jong, J.E. van Timmeren, A. Ibrahim, I. Compter, J. Peerlings, S. Sanduleanu, T. Refaee, S. Keek, R.T.H.M. Larue, Y. van Wijk, A.J.G. Even, A. Jochems, M.S. Barakat, R.T.H. Leijenaar, P. Lambin, Decision Support Systems in Oncology, *JCO Clin Cancer Inform.* 3 (2019) 1–9. <https://doi.org/10.1200/CCI.18.00001>.
2. P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J.W.L. Aerts, Radiomics: extracting more information from medical images using advanced feature analysis, *Eur. J. Cancer.* 48 (2012) 441–446. <https://doi.org/10.1016/j.ejca.2011.11.036>.
3. R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images Are More than Pictures, They Are Data, *Radiology.* 278 (2016) 563–577. <https://doi.org/10.1148/radiol.2015151169>.
4. P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, *Nat. Rev. Clin. Oncol.* 14 (2017) 749–762. <https://doi.org/10.1038/nrclinonc.2017.141>.
5. T. Refaee, G. Wu, A. Ibrahim, I. Halilaj, R.T.H. Leijenaar, W. Rogers, H.A. Gietema, L.E.L. Hendriks, P. Lambin, H.C. Woodruff, The Emerging Role of Radiomics in COPD and Lung Cancer, *Respiration.* 99 (2020) 99–107. <https://doi.org/10.1159/000505429>.
6. R.C. Hardie, S.K. Rogers, T. Wilson, A. Rogers, Performance analysis of a new computer aided detection system for identifying lung nodules on chest radiographs, *Med. Image Anal.* 12 (2008) 240–258. <https://doi.org/10.1016/j.media.2007.10.004>.
7. G.J.R. Cook, M. Siddique, B.P. Taylor, C. Yip, S. Chicklore, V. Goh, Radiomics in PET: principles and applications, *Clinical and Translational Imaging.* 2 (2014) 269–276. <https://doi.org/10.1007/s40336-014-0064-0>.
8. S. Chicklore, V. Goh, M. Siddique, A. Roy, P.K. Marsden, G.J.R. Cook, Quantifying tumour heterogeneity in 18F-FDG PET/CT imaging by texture analysis, *Eur. J. Nucl. Med. Mol. Imaging.* 40 (2013) 133–140. <https://doi.org/10.1007/s00259-012-2247-0>.
9. C. Swanton, Intratumor heterogeneity: evolution through space and time, *Cancer Res.* 72 (2012) 4875–4882. <https://doi.org/10.1158/0008-5472.CAN-12-2217>.
10. M. Gerlinger, A.J. Rowan, S. Horswell, M. Math, J. Larkin, D. Endesfelder, E. Gronroos, P. Martinez, N. Matthews, A. Stewart, P. Tarpey, I. Varela, B. Phillimore, S. Begum, N.Q. McDonald, A. Butler, D. Jones, K. Raine, C. Latimer, C.R. Santos, M. Nohadani, A.C. Eklund, B. Spencer-Dene, G. Clark, L. Pickering, G. Stamp, M. Gore, Z. Szallasi, J. Downward, P.A. Futreal, C. Swanton, Intratumor heterogeneity and branched evolution revealed by multiregion sequencing, *N. Engl. J. Med.* 366 (2012) 883–892. <https://doi.org/10.1056/NEJMoa1113205>.
11. T.M. Soo, M. Bernstein, J. Provias, R. Tasker, A. Lozano, A. Guha, Failed stereotactic biopsy in a series of 518 cases, *Stereotact. Funct. Neurosurg.* 64 (1995) 183–196. <https://doi.org/10.1159/000098747>.

Chapter 2

12. S.S.F. Yip, H.J.W.L. Aerts, Applications and limitations of radiomics, *Phys. Med. Biol.* 61 (2016) R150–66. <https://doi.org/10.1088/0031-9155/61/13/R150>.
13. Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature*. 521 (2015) 436–444. <https://doi.org/10.1038/nature14539>.
14. W.S. McCulloch, W. Pitts, A logical calculus of the ideas immanent in nervous activity, *Bull. Math. Biophys.* 5 (1943) 115–133. <https://doi.org/10.1007/BF02478259>.
15. L. Hongtao, Z. Qinchuan, Applications of Deep Convolutional Neural Network in Computer Vision, *J. Data Acquisition Process.* (2016). http://en.cnki.com.cn/Article_en/CJFDTotal-SJCJ201601001.htm.
16. H. Shirani-Mehr, Applications of deep learning to sentiment analysis of movie reviews, *Tech. Rep. NAVTRADEVCEEN.* (2014) 1–8. <https://pdfs.semanticscholar.org/59c8/643ca06357da0c8977fe4ea757119ba98003.pdf>.
17. L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, *APSIPA Transactions on Signal and Information Processing.* 3 (2014). <https://doi.org/10.1017/atsip.2013.9>.
18. K. Janocha, W.M. Czarnecki, On Loss Functions for Deep Neural Networks in Classification, *Schedae Informaticae.* 1/2016 (2017). <https://doi.org/10.4467/20838476si.16.004.6185>.
19. A. Voulodimos, N. Doulamis, A. Doulamis, E. Protopapadakis, Deep Learning for Computer Vision: A Brief Review, *Comput. Intell. Neurosci.* 2018 (2018) 7068349. <https://doi.org/10.1155/2018/7068349>.
20. R. Simhambhatla, K. Okiah, S. Kuchkula, R. Slater, Self-Driving Cars: Evaluation of Deep Learning Techniques for Object Detection in Different Driving Conditions, *SMU Data Science Review.* 2 (2019) 23. <https://scholar.smu.edu/datasciencereview/vol2/iss1/23/> (accessed May 14, 2020).
21. D. Shen, G. Wu, H.-I. Suk, Deep Learning in Medical Image Analysis, *Annu. Rev. Biomed. Eng.* 19 (2017) 221–248. <https://doi.org/10.1146/annurev-bioeng-071516-044442>.
22. A. Ibrahim, M. Vallières, H. Woodruff, S. Primakov, M. Beheshti, S. Keek, T. Refaee, S. Sanduleanu, S. Walsh, O. Morin, P. Lambin, R. Hustinx, F.M. Mottaghy, Radiomics Analysis for Clinical Decision Support in Nuclear Medicine, *Semin. Nucl. Med.* 49 (2019) 438–449. <https://doi.org/10.1053/j.semnuclmed.2019.06.005>.
23. L.R. Cardon, H. Watkins, Waiting for the working draft from the human genome project. A huge achievement, but not of immediate medical use, *BMJ.* 320 (2000) 1223–1224. <https://doi.org/10.1136/bmj.320.7244.1223>.
24. N.J. Schork, Personalized medicine: Time for one-person trials, *Nature*. 520 (2015) 609–611. <https://doi.org/10.1038/520609a>.
25. C. Parmar, J.D. Barry, A. Hosny, J. Quackenbush, H.J.W.L. Aerts, Data Analysis Strategies in Medical Imaging, *Clin. Cancer Res.* 24 (2018) 3492–3499. <https://doi.org/10.1158/1078-0432.CCR-18-0385>.
26. J. Wang, L. Shen, H. Zhong, Z. Zhou, P. Hu, J. Gan, R. Luo, W. Hu, Z. Zhang, Radiomics features on radiotherapy treatment planning CT can predict patient survival in locally advanced rectal

- cancer patients, *Sci. Rep.* 9 (2019) 15346. <https://doi.org/10.1038/s41598-019-51629-4>.
27. S. Bae, Y.S. Choi, S.S. Ahn, J.H. Chang, S.-G. Kang, E.H. Kim, S.H. Kim, S.-K. Lee, Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction, *Radiology.* 289 (2018) 797–806. <https://doi.org/10.1148/radiol.2018180200>.
 28. A. Oikonomou, F. Khalvati, P.N. Tyrrell, M.A. Haider, U. Tarique, L. Jimenez-Juan, M.C. Tjong, I. Poon, A. Eilaghi, L. Ehrlich, P. Cheung, Radiomics analysis at PET/CT contributes to prognosis of recurrence and survival in lung cancer treated with stereotactic body radiotherapy, *Sci. Rep.* 8 (2018) 4003. <https://doi.org/10.1038/s41598-018-22357-y>.
 29. M. Kirienko, L. Cozzi, L. Antunovic, L. Lozza, A. Fogliata, E. Voulaz, A. Rossi, A. Chiti, M. Sollini, Prediction of disease-free survival by the PET/CT radiomic signature in non-small cell lung cancer patients undergoing surgery, *Eur. J. Nucl. Med. Mol. Imaging.* 45 (2018) 207–217. <https://doi.org/10.1007/s00259-017-3837-7>.
 30. P. Kickingereder, S. Burth, A. Wick, M. Götz, O. Eidel, H.-P. Schlemmer, K.H. Maier-Hein, W. Wick, M. Bendszus, A. Radbruch, D. Bonekamp, Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models, *Radiology.* 280 (2016) 880–889. <https://doi.org/10.1148/radiol.2016160845>.
 31. W. Wu, C. Parmar, P. Grossmann, J. Quackenbush, P. Lambin, J. Bussink, R. Mak, H.J.W.L. Aerts, Exploratory Study to Identify Radiomics Classifiers for Lung Cancer Histology, *Front. Oncol.* 6 (2016) 71. <https://doi.org/10.3389/fonc.2016.00071>.
 32. M. Wu, H. Tan, F. Gao, J. Hai, P. Ning, J. Chen, S. Zhu, M. Wang, S. Dou, D. Shi, Predicting the grade of hepatocellular carcinoma based on non-contrast-enhanced MRI radiomics signature, *Eur. Radiol.* 29 (2019) 2802–2811. <https://doi.org/10.1007/s00330-018-5787-2>.
 33. M. Vallières, C.R. Freeman, S.R. Skamene, I. El Naqa, A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities, *Phys. Med. Biol.* 60 (2015) 5471–5496. <https://doi.org/10.1088/0031-9155/60/14/5471>.
 34. S. Trebeschi, S. G. Drago, N. J. Birkbak, I. Kurilova, A. M. Calin, A. Delli Pizzi, F. Lalezari, D. M. J. Lambregts, M. W. Rohaan, C. Parmar, E. A. Rozeman, K. J. Hartemink, C. Swanton, J. B. A. G. Haanen, C. U. Blank, E. F. Smit, R. G. H. Beets-Tan & H. J. W. L. Aerts, Predicting response to cancer immunotherapy using noninvasive radiomic biomarkers, *Annals of.* (2019). <https://academic.oup.com/annonc/article-abstract/30/6/998/5416144>.
 35. N. Horvat, H. Veeraraghavan, M. Khan, I. Blazic, J. Zheng, M. Capanu, E. Sala, J. Garcia-Aguilar, M.J. Gollub, I. Petkovska, MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy, *Radiology.* 287 (2018) 833–843. <https://doi.org/10.1148/radiol.2018172300>.
 36. J. Alirezaie, M.E. Jernigan, C. Nahmias, Automatic segmentation of cerebral MR images using artificial neural networks, *IEEE Trans. Nucl. Sci.* 45 (1998) 2174–2182. <https://doi.org/10.1109/23.708336>.
 37. J. Jiang, Y.-C. Hu, C.-J. Liu, D. Halpenny, M.D. Hellmann, J.O. Deasy, G. Mageras, H. Veeraraghavan, Multiple Resolution Residually Connected Feature Streams for Automatic Lung

- Tumor Segmentation From CT Images, *IEEE Trans. Med. Imaging.* 38 (2019) 134–144. <https://doi.org/10.1109/TMI.2018.2857800>.
38. D. Yi, M. Zhou, Z. Chen, O. Gevaert, 3-D Convolutional Neural Networks for Glioblastoma Segmentation, *arXiv [cs.CV]*. (2016). <http://arxiv.org/abs/1611.04534>.
 39. B.H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B.B. Avants, N. Ayache, P. Buendía, D.L. Collins, N. Cordier, J.J. Corso, A. Criminisi, T. Das, H. Delingette, Ç. Demiralp, C.R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K.M. Iftekharuddin, R. Jena, N.M. John, E. Konukoglu, D. Lashkari, J.A. Mariz, R. Meier, S. Pereira, D. Precup, S.J. Price, T.R. Raviv, S.M.S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C.A. Silva, N. Sousa, N.K. Subbanna, G. Szekely, T.J. Taylor, O.M. Thomas, N.J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D.H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, K. Van Leemput, The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), *IEEE Trans. Med. Imaging.* 34 (2015) 1993–2024. <https://doi.org/10.1109/TMI.2014.2377694>.
 40. L. Chen, C. Shen, S. Li, G. Maquilan, K. Albuquerque, M.R. Folkert, J. Wang, Automatic PET cervical tumor segmentation by deep learning with prior information, in: *Medical Imaging 2018: Image Processing*, International Society for Optics and Photonics, 2018: p. 1057436. <https://doi.org/10.1117/12.2293926>.
 41. D. Ardila, A.P. Kiraly, S. Bharadwaj, B. Choi, J.J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D.P. Naidich, S. Shetty, Author Correction: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, *Nat. Med.* 25 (2019) 1319. <https://doi.org/10.1038/s41591-019-0536-x>.
 42. S.A. Abdelaziz Ismael, A. Mohammed, H. Hefny, An enhanced deep learning approach for brain cancer MRI images classification using residual networks, *Artif. Intell. Med.* 102 (2020) 101779. <https://doi.org/10.1016/j.artmed.2019.101779>.
 43. L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, M. Schäfers, 18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks, *Radiology.* 294 (2020) 445–452. <https://doi.org/10.1148/radiol.2019191114>.
 44. S.L.F. Walsh, L. Calandriello, M. Silva, N. Sverzellati, Deep learning for classifying fibrotic lung disease on high-resolution computed tomography: a case-cohort study, *Lancet Respir Med.* 6 (2018) 837–845. [https://doi.org/10.1016/S2213-2600\(18\)30286-8](https://doi.org/10.1016/S2213-2600(18)30286-8).
 45. Y. Ding, J.H. Sohn, M.G. Kawczynski, H. Trivedi, R. Harnish, N.W. Jenkins, D. Lituiev, T.P. Copeland, M.S. Aboian, C. Mari Aparici, S.C. Behr, R.R. Flavell, S.-Y. Huang, K.A. Zalocusky, L. Nardo, Y. Seo, R.A. Hawkins, M. Hernandez Pampaloni, D. Hadley, B.L. Franc, A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18F-FDG PET of the Brain, *Radiology.* 290 (2019) 456–464. <https://doi.org/10.1148/radiol.2018180958>.
 46. K. Oh, Y.-C. Chung, K.W. Kim, W.-S. Kim, I.-S. Oh, Author Correction: Classification and Visualization of Alzheimer’s Disease using Volumetric Convolutional Neural Network and Transfer

- Learning, *Sci. Rep.* 10 (2020) 5663. <https://doi.org/10.1038/s41598-020-62490-1>.
47. B. Lou, S. Doken, T. Zhuang, D. Wingerter, M. Gidwani, N. Mistry, L. Ladic, A. Kamen, M.E. Abazeed, An image-based deep learning framework for individualising radiotherapy dose: a retrospective analysis of outcome prediction, *The Lancet Digital Health.* 1 (2019) e136–e147. [https://doi.org/10.1016/s2589-7500\(19\)30058-5](https://doi.org/10.1016/s2589-7500(19)30058-5).
 48. P.-P. Ypsilantis, M. Siddique, H.-M. Sohn, A. Davies, G. Cook, V. Goh, G. Montana, Predicting Response to Neoadjuvant Chemotherapy with PET Imaging Using Convolutional Neural Networks, *PLoS One.* 10 (2015) e0137036. <https://doi.org/10.1371/journal.pone.0137036>.
 49. K. Strimbu, J.A. Tavel, What are biomarkers?, *Curr. Opin. HIV AIDS.* 5 (2010) 463. <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3078627/>.
 50. H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haihe-Kains, D. Rietveld, F. Hoebbers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, *Nat. Commun.* 5 (2014) 4006. <https://doi.org/10.1038/ncomms5006>.
 51. U. Schneider, E. Pedroni, A. Lomax, The calibration of CT Hounsfield units for radiotherapy treatment planning, *Phys. Med. Biol.* 41 (1996) 111–124. <https://doi.org/10.1088/0031-9155/41/1/009>.
 52. L.G. Nyúl, J.K. Udupa, On standardizing the MR image intensity scale, *Magn. Reson. Med.* 42 (1999) 1072–1081. [>10.1002/\(sici\)1522-2594\(199912\)42:6<1072::aid-mrm11>3.0.co;2-m.](https://doi.org/3.0.co;2-m.)
 53. R. Berenguer, M.D.R. Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M.V. Villas, F. Mansilla Legorburo, S. Sabater, Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters, *Radiology.* 288 (2018) 407–415. <https://doi.org/10.1148/radiol.2018172361>.
 54. J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific?, *Tomography.* 2 (2016) 361–365. <https://doi.org/10.18383/j.tom.2016.00208>.
 55. L. Lu, R.C. Ehmke, L.H. Schwartz, B. Zhao, Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings, *PLoS One.* 11 (2016) e0166550. <https://doi.org/10.1371/journal.pone.0166550>.
 56. D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring Computed Tomography Scanner Variability of Radiomics Features, *Invest. Radiol.* 50 (2015) 757–765. <https://doi.org/10.1097/RLI.0000000000000180>.
 57. I. Zhovannik, J. Bussink, A. Traverso, Z. Shi, P. Kalendralis, L. Wee, A. Dekker, R. Fijten, R. Monshouwer, Learning from scanners: Bias reduction and feature correction in radiomics, *Clin Transl Radiat Oncol.* 19 (2019) 33–38. <https://doi.org/10.1016/j.ctro.2019.07.003>.
 58. A. Webb, G.C. Kagadis, Introduction to Biomedical Imaging, *Med. Phys.* 30 (2003) 2267–2267. <https://doi.org/10.1118/1.1589017>.
 59. S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles, A. Traverso,

- D. Jaffray, U. Metsler, J. Xie, K. Han, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, *Radiother. Oncol.* 135 (2019) 107–114. <https://doi.org/10.1016/j.radonc.2019.03.001>.
60. J. Peerlings, H.C. Woodruff, J.M. Winfield, A. Ibrahim, B.E. Van Beers, A. Heerschap, A. Jackson, J.E. Wildberger, F.M. Mottaghy, N.M. DeSouza, P. Lambin, Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial, *Sci. Rep.* 9 (2019) 4800. <https://doi.org/10.1038/s41598-019-41344-5>.
61. K.T. Bae, Intravenous contrast medium administration and scan timing at CT: considerations and approaches, *Radiology.* 256 (2010) 32–61. <https://doi.org/10.1148/radiol.10090908>.
62. G.J.R. Cook, G. Azad, K. Owczarczyk, M. Siddique, V. Goh, Challenges and Promises of PET Radiomics, *Int. J. Radiat. Oncol. Biol. Phys.* 102 (2018) 1083–1089. <https://doi.org/10.1016/j.ijrobp.2017.12.268>.
63. F. Tixier, M. Hatt, C.C. Le Rest, A. Le Pogam, L. Corcos, D. Visvikis, Reproducibility of tumor uptake heterogeneity characterization through textural feature analysis in 18F-FDG PET, *J. Nucl. Med.* 53 (2012) 693–700. <https://doi.org/10.2967/jnumed.111.099127>.
64. M. Hatt, F. Tixier, C.C. Le Rest, O. Pradier, Robustness of intratumour 18F-FDG PET uptake heterogeneity quantification for therapy response prediction in oesophageal carcinoma, *European Journal of.* (2013). <https://link.springer.com/article/10.1007/s00259-013-2486-8>.
65. R.T.H. Leijenaar, G. Nalbantov, S. Carvalho, W.J.C. van Elmpt, E.G.C. Troost, R. Boellaard, H.J.W.L. Aerts, R.J. Gillies, P. Lambin, The effect of SUV discretization in quantitative FDG-PET Radiomics: the need for standardized methodology in tumor texture analysis, *Sci. Rep.* 5 (2015) 11075. <https://doi.org/10.1038/srep11075>.
66. B.A. Altazi, G.G. Zhang, D.C. Fernandez, M.E. Montejo, D. Hunt, J. Werner, M.C. Biagioli, E.G. Moros, Reproducibility of F18-FDG PET radiomic features for different cervical tumor segmentation methods, gray-level discretization, and reconstruction algorithms, *J. Appl. Clin. Med. Phys.* 18 (2017) 32–48. <https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12170>.
67. I. Shiri, A. Rahmim, P. Ghaffarian, P. Geramifar, H. Abdollahi, A. Bitarafan-Rajabi, The impact of image reconstruction settings on 18F-FDG PET radiomic features: multi-scanner phantom and patient studies, *Eur. Radiol.* 27 (2017) 4498–4509. <https://doi.org/10.1007/s00330-017-4859-z>.
68. R.T.H. Leijenaar, S. Carvalho, E.R. Velazquez, W.J.C. van Elmpt, C. Parmar, O.S. Hoekstra, C.J. Hoekstra, R. Boellaard, A.L.A.J. Dekker, R.J. Gillies, H.J.W.L. Aerts, P. Lambin, Stability of FDG-PET Radiomics features: an integrated analysis of test-retest and inter-observer variability, *Acta Oncol.* 52 (2013) 1391–1397. <https://doi.org/10.3109/0284186X.2013.812798>.
69. M. Pavic, M. Bogowicz, X. Würms, S. Glatz, T. Finazzi, O. Riesterer, J. Roesch, L. Rudofsky, M. Friess, P. Veit-Haibach, M. Huellner, I. Opitz, W. Weder, T. Frauenfelder, M. Guckenberger, S. Tanadini-Lang, Influence of inter-observer delineation variability on radiomics stability in different tumor sites, *Acta Oncol.* 57 (2018) 1070–1074. <https://doi.org/10.1080/0284186X.2018.1445283>.
70. M. Hatt, M. Vallieres, D. Visvikis, A. Zwanenburg, IBSI: an international community radiomics standardization initiative, *J. Nucl. Med.* 59 (2018) 287–287. <http://jnm.snmjournals.org/>

content/59/supplement_1/287.abstract.

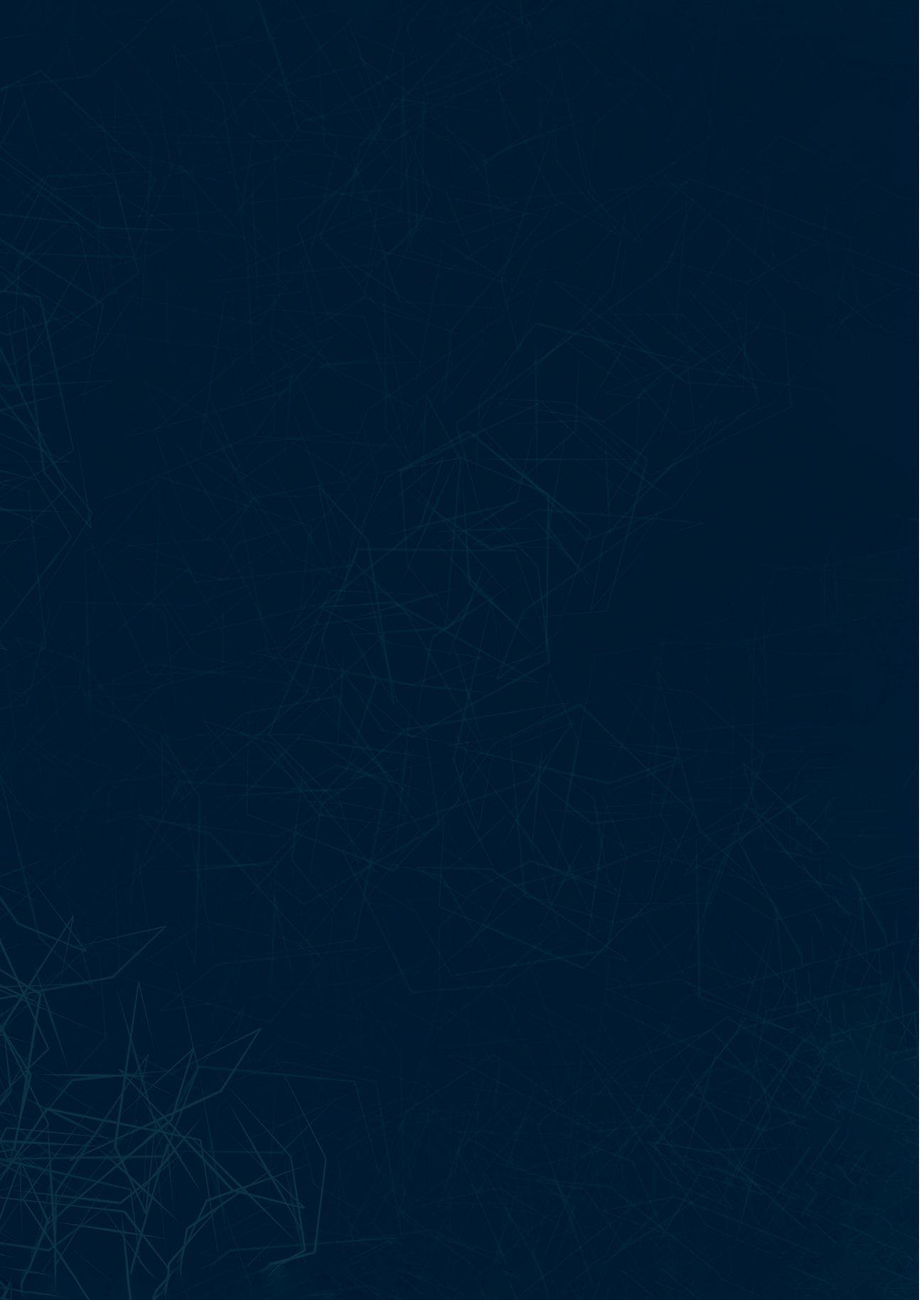
71. A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desserot, N. Dinapoli, C.V. Dinh, S. EcheGARAY, I. El Naqa, A.Y. Fedorov, R. Gatta, R.J. Gillies, V. Goh, M. Götz, M. Guckenberger, S.M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R.T.H. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K.H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E.A.G. Pfaehler, A. Rahmim, A.U.K. Rao, J. Scherer, M.M. Siddique, N.M. Sijtsema, J. Socarras Fernandez, E. Spezi, R.J.H.M. Steenbakkens, S. Tanadini-Lang, D. Thorwarth, E.G.C. Troost, T. Upadhaya, V. Valentini, L.V. van Dijk, J. van Griethuysen, F.H.P. van Velden, P. Whybra, C. Richter, S. Löck, The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping, *Radiology*. (2020) 191145. <https://doi.org/10.1148/radiol.2020191145>.
72. W.E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods, *Biostatistics*. 8 (2007) 118–127. <https://doi.org/10.1093/biostatistics/kxj037>.
73. F. Orlhac, F. Frouin, C. Nioche, N. Ayache, I. Buvat, Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics, *Radiology*. 291 (2019) 53–59. <https://doi.org/10.1148/radiol.2019182023>.
74. F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, I. Buvat, A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET, *J. Nucl. Med.* 59 (2018) 1321–1328. <https://doi.org/10.2967/jnumed.117.199935>.

Acknowledgments

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015, n° 694812 - Hypoximmuno), ERC-2018-PoC (n° 81320 – CL-IO). We further acknowledge the financial support from Maastricht-Liege imaging valley grant. This research is also supported by the Dutch technology Foundation STW (grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL - n°673780), EUROSTARS (DART, DECIDE), the European Program H2020-2015-17 (BD2Decide - PHC30-689715, ImmunoSABR - n° 733008, PREDICT - ITN - n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 'CLEARLY'- n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine ('Euradiomics'). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2.

Competing Interests

Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr. Lambin has shares in the company Oncoradiomics SA and Convert pharmaceuticals SA and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patentable invention (softwares) licensed to ptTheragnostic/DNAmito, Oncoradiomics and Health Innovation Ventures. Dr. Woodruff has (minority) shares in the company Oncoradiomics.



PART II

A large, white, stylized number 3 is centered on a blue, textured, watercolor-like background. The background consists of various shades of blue, from light to dark, with a mottled, organic appearance. The number 3 is a simple, clean, sans-serif font. The overall composition is abstract and artistic.

Chapter 3

MRI-Based Radiomics Analysis for the Pretreatment Prediction of Pathologic Complete Tumor Response to Neoadjuvant Systemic Therapy in Breast Cancer Patients: A Multicenter Study

Authors

Renée W. Y. Granzier, Abdalla Ibrahim, Sergey P. Primakov,
Sanaz Samiei, Thiemo J. A. van Nijnatten, Maaïke de Boer, Esther M. Heuts,
Frans-Jan Hulsmans, Avishek Chatterjee, Philippe Lambin,
Marc B. I. Lobbes, Henry C. Woodruff and Marjolein L. Smidt

Adapted from

Cancers. 2021 Jan;13(10):2447

DOI

10.3390/cancers13102447

Abstract

This retrospective study investigated the value of pretreatment contrast-enhanced Magnetic Resonance Imaging (MRI)-based radiomics for the prediction of pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients. A total of 292 breast cancer patients, with 320 tumors, who were treated with neo-adjuvant systemic therapy and underwent a pretreatment MRI exam were enrolled. As the data were collected in two different hospitals with five different MRI scanners and varying acquisition protocols, three different strategies to split training and validation datasets were used. Radiomics, clinical, and combined models were developed using random forest classifiers in each strategy. The analysis of radiomics features had no added value in predicting pathologic complete tumor response to neoadjuvant systemic therapy in breast cancer patients compared with the clinical models, nor did the combined models perform significantly better than the clinical models. Further, the radiomics features selected for the models and their performance differed with and within the different strategies. Due to previous and current work, we tentatively attribute the lack of improvement in clinical models following the addition of radiomics to the effects of variations in acquisition and reconstruction parameters. The lack of reproducibility data (i.e., test-retest or similar) meant that this effect could not be analyzed. These results indicate the need for reproducibility studies to preselect reproducible features in order to properly assess the potential of radiomics.

Keywords

breast cancer; MRI; neoadjuvant systemic therapy; response prediction; radiomics

Introduction

Neoadjuvant systemic therapy (NST) is increasingly administered in the treatment of breast cancer. The number of breast cancer patients receiving NST varies between 17% and 70% and depends mainly on breast cancer subtype and tumor size ^[1,2]. NST allows monitoring of *in vivo* tumor response, potentially decreasing tumor size and thus enabling breast-conserving surgery ^[1,3,4]. Unfortunately, not all patients respond well to NST, with tumor response ranging from pathologic complete tumor response (pCR) to non-response and sometimes even progression of disease. Predicting which patients will respond well to NST and achieve tumor pCR could lead to modifications of treatment plans. In current clinical practice, magnetic resonance imaging (MRI) assessment combined with clinical (tumor) characteristics is used to determine tumor response to NST ^[5-7]. However, the diagnostic accuracy of the MRI with regard to tumor response evaluation is insufficiently accurate (76.1%) to adapt clinical treatment plans ^[8]. Furthermore, two studies investigated the use of ultrasound-guided biopsies to identify pCR after NST ^[9,10]. Unfortunately, the results showed that these biopsies are not accurate enough to identify pCR that surgery can be omitted ^[11].

Radiomics, a quantitative image analysis technique, could play a role predicting pCR from pretreatment dynamic contrast-enhanced (DCE)-MRI exams. Radiomics extracts large amounts of quantitative features from medical imaging, including MRI. These features capture information on the underlying heterogeneous structure of the region of interest (ROI), describing volume and shape, intensities and textures ^[12]. Radiomics' non-invasive ability to characterize the three-dimensional ROI, combined with the availability of ever-growing amounts of (longitudinal) imaging data and its cost-effectiveness, all contribute to the potential use of radiomics in personalized medicine ^[13-16]. The emergence of radiomics has so far mainly been applied in the field of clinical oncology and has also permeated breast cancer research.

Several MRI-based radiomics studies have reported promising results regarding the prediction of pCR to NST in breast cancer patients based on pretreatment scans ^[17-21]. However, the evidence from these studies is limited due to the relatively small sample sizes ranging from 29 to 100 patients and the lack of external validation datasets. Despite the promising potential of radiomics, several hurdles that impede the clinical implementation of radiomics models have been identified. One of these is the sensitivity of radiomics features to the variations in acquisition and reconstruction parameters across different imaging modalities ^[22-26], and some features were found not to be reproducible even in test-retest scenarios ^[27-29].

This study aimed to investigate the potential of pretreatment contrast-enhanced MRI-

based radiomics for the prediction of pCR to NST in breast cancer patients. We hypothesized that radiomics models trained and validated on data from two independent cohorts could add information to the prediction of tumor response to NST and that combined with clinical models can improve prediction accuracy. During our analysis, the sensitivity of radiomics features to the variations in acquisition and reconstruction parameters was established.

Materials and Methods

Study Population

In this multicenter study, imaging, and clinical data from consecutive women with histopathologically confirmed invasive breast cancer were retrospectively collected from two hospitals in the Netherlands (MUMC+—Maastricht University Medical Center and ZMC—Zuyderland Medical Center) between January 2011 and December 2018. The inclusion criteria were as follows: (i) treated with NST, (ii) have undergone pretreatment DCE-MRI in one of the two participating hospitals, and (iii) breast surgery after NST with histopathological outcome. Exclusion criteria were as follows: (i) histopathologically confirmed inflammatory breast cancer without the possibility of unequivocal tumor segmentation, (ii) MRI exam artefacts, if also rejected for visual assessment by the breast radiologist, (iii) non-standard chemotherapy regimens, deviating from the Dutch breast cancer guidelines, (iv) unfinished NST, and (v) no access to the patient's medical record. In the case of multifocal breast cancer, all histopathologically confirmed invasive tumors were included in the study. The institutional research board of both hospitals approved the study and waived the requirement for informed consent.

Study Strategy

As different MRI scanners with varying acquisition and reconstruction parameters were used in the two hospitals, it was decided to develop separate prediction models (radiomics, clinical, and a combination of the two) for both cohorts and to validate them on each other (strategies 1 and 2). Therefore, all feature reduction, selection, and modeling procedures were performed on both data cohorts. A third modelling strategy was based on a mixture of both datasets divided into 70% training and 30% validation cohort. Feature selection and model building was performed on 70% of the training data and tested on the remaining 30% of the training data. The process of splitting the data into training and testing was iterated 100 times, maintaining class imbalance and ensuring that tumors from one patient were selected either in the training data or in the testing data. Figure 1A shows an overview of the selected data per strategy.

Clinical and Pathological Data

Clinical and pathological data were retrieved from patients' medical records and included

age, clinical and pathological tumor, nodes, and metastases (TNM) stage, tumor grade, tumor histology, breast cancer subtype, and NST regimen. The majority of patients were treated with an anthracycline- and taxane-based NST regimen; the remaining received a taxane-based only NST regimen. Human epidermal growth factor receptor 2 (HER2) positive tumors received additional treatment with trastuzumab and/or pertuzumab. After completion of NST, all patients underwent breast surgery. The surgical specimens of all patients were evaluated via standard histopathological analysis by breast pathologists in the two participating hospitals. The breast tumor response was assessed by the Miller–Payne or Pinder grading systems [30,31]. In this study, tumors were defined as pCR when classified as grade 5 using the Miller–Payne classification or classified as 1i and 1ii using the Pinder classification (pCR; ductal carcinoma in situ may be present).

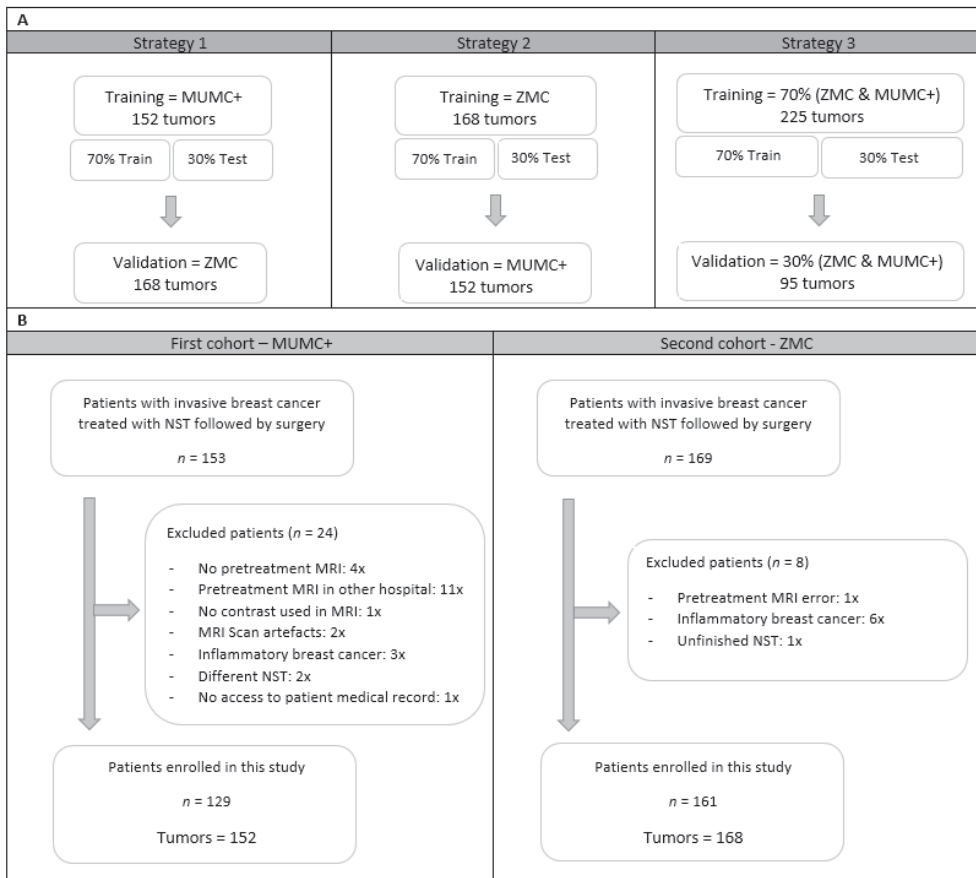


Figure I: An overview of training, test, and validation data cohorts for the three strategies (A) and a flowchart from patient selection for the two different hospitals (B). Abbreviations, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center, NST = Neoadjuvant Systemic Therapy, MRI = Magnetic Resonance Imaging.

Table I: Scanning Parameters.

Hospital	Scanner	Total MRI Exam No.	Group	No. of Tumors for Specific Scanning Parameters	Pixel Spacing	Acquisition Matrix (n)	Slice Thickness (mm)	TR/TE (ms) (n)	Spacing between Slices	Flip Angle
MUMC+	Philips 1.5T (Ingenia)	124	a	44	(0.97, 0.97)	340 × 340	1	3.4/7.5 3.5/7.6	1	10°
			b	66	(0.95, 0.95)	378 × 314 (28) 380 × 318 (23) 380 × 316 (18)	1	3.2/7.1 3.4/7.5 3.5/7.6	1	10°
			c	9	(0.80, 0.80)	344 × 344	1	3.4/7.5	1	10°
			d	3	(0.92, 0.92)	400 × 333 (2) 398 × 331 (1)	1	3.5/7.6 3.4/7.5	1	10°
			e	1	(0.88, 0.88)	384 × 368	1	3.4/7.5	1	10°
			f	1	(0.85, 0.85)	384 × 278	1	2.9/6.5	1	10°
	Philips 1.5T (Intera)	28	a	25	(0.97, 0.97)	340 × 337	1	3.4/7.4-7.6	1	10°
			b	1	(0.99, 0.99)	376 × 376	1	3.4/7.4	1	10°
			c	1	(0.95, 0.95)	364 × 364	1	3.4/7.5	1	10°
			d	1	(0.85, 0.85)	368 × 368	1	3.4/7.4	1	10°
			a	94	(0.97, 0.97)	340 × 338	2	3.4/6.9-7.0	1	12°
			b	28	(0.96, 0.96)	372 × 368 (15) 372 × 370 (13)	2	3.4/6.9-7.0	1	12°
ZMC	Siemens 3.0T (Skyra)	39	c	1	(0.90, 0.90)	392 × 388	2	3.4/6.9	1	12°
			a	39	(0.69, 0.69)	288 × 288	2	1.2/4.0	unknown	10°
			a	6	(0.89, 0.89)	224 × 202	2	2.4/6.1	unknown	10°
	Siemens 1.5T (Avanto_fit)	6								

Abbreviations, MRI = Magnetic Resonance Imaging, TR = Repetition Time, TE = Echo Time, T = Tesla, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center.

Imaging Data

For all patients, pretreatment MRI exams were collected containing fat-suppressed 3D THRIVE DCE T1-weighted (T1W), T2-weighted in the MUMC and fat-suppressed T2-weighted in the ZMC, and diffusion weighted imaging sequences. It was decided to only use the peak-enhanced phase of the DCE-T1W images for the radiomics analysis as tumors are best visible on this sequence [32,33]. The DCE-T1W images were obtained before and after intravenous injection of gadolinium-based contrast Gadobutrol (GadovistTM (EU)) with a volume of 15 mL and a flow rate of 2 mL/sec. A 105 s temporal resolution protocol was used in the MUMC+ and a 20 s temporal resolution protocol in the ZMC, resulting in five and nineteen post-contrast images for each patient in the MUMC+ and ZMC, respectively. Images were acquired using 1.5T (Ingenia, Intera, and Achieva by Philips Medical system and Avanto Fit by Siemens) and 3.0T (Skyra by Siemens) MRI scanners. All patients were scanned in prone-position using a dedicated breast-coil. DCE-T1W MRI acquisition protocols from both hospitals can be found in Table 1. Sequence parameters varied per MRI scanner and hospital, reflecting the heterogeneity in medical images used in daily clinical practice.

Tumor Segmentation

The images acquired at tumor peak enhancement, at approximately two minutes' post-contrast administration, were used for the 3D ROI segmentation and further radiomics analysis, as tumors are best assessed on these images. All histologically confirmed invasive tumors were segmented manually using Mirada Medical's DBx 1.2.0.59 (64-bit, Oxford, UK) software by a medical researcher with three years of experience (RG), supervised by a dedicated breast radiologist with 14 years of experience (ML). During segmentation, the radiology reports were accessible, and adjustment of image grayscale was allowed to optimize the visualization of the tumor. To gauge any bias introduced by inter-observer segmentation variability, 129 tumors from 102 patients acquired at MUMC+ were segmented by four observers independently with different degrees of experience in breast MR imaging (RG, ML, resident with three years of MRI experience (TvN), and a medical student with no experience (NV)) [34].

Image Pre-Processing and Feature Selection

Image pre-processing of the two-minute postcontrast-T1W images was performed after tumor segmentation using an in-house developed pipeline and using a widely used proposed pre-processing method by Pyradiomics [35,36]. The in-house developed pipeline started first by applying bias field correction to every image using MIM software (version 6.9.4, Cleveland, Ohio, Unites States) to correct for nonuniform grayscale intensities in the MRI caused by field inhomogeneities. Second, in order to minimize acquisition-related radiomics variability, voxel dimensions were standardized across the cohorts to arrive at an isotropic voxel resolution of 1 mm³ by means of cubic

interpolation^[37]. Third, to homogenize arbitrary MRI units and clip image intensities to a certain range, a histogram matching technique was applied, adjusting the pixel values of the MR image such that its histogram matched that of the target MR image from the training data cohort^[38–40]. Further gray value filtering was applied to generate MRIs with comparable gray value range and to enhance the contrast of the image using the following filtering parameters: window level (WL: 3050) and window width (WW: 2950). Filtering parameters were found when exploring the images after the histogram matching step. Fourth, to reduce high frequency noise and optimize handling of the image, grayscale values were resampled using a fixed bin width of 24, which reduced both image noise and computation times when extracting radiomics features from the ROI^[41]. The pre-processing method proposed by Pyradiomics was applied after images' bias field correction and consisted of z-score normalization, resampling to isotropic voxel resolution of 1 mm³, and image discretization using a bin width of 100 to reach an ideal number of bins between 16 and 128^[12].

For each ROI, 833 features were extracted using the Pyradiomics software (version 3.0). The extracted radiomics features included first-order statistics features (18), shape-based features (14), gray-level co-occurrence matrix features (GLCM) (22), gray-level run length matrix features (GLRLM) (16), gray-level size zone matrix features (GLSZM) (16), neighboring gray tone difference matrix features (NGTDM) (5), and gray-level dependence matrix features (GLDM) (14) from both unfiltered and filtered (eight wavelet decompositions) images.

Feature Selection and Radiomics Model Development

All feature selection steps followed by model development were performed on the 70% training data for each iteration. First, features sensitive to interobserver segmentation variabilities were removed using an intraclass correlation coefficient (ICC) cut-off value >0.75 (29). Consecutively, features with zero or small variance (with the frequency ratio between the most common value and the second most common value larger than 95/5) were removed. This was followed by the removal of highly correlated features using pairwise Spearman correlation ($|r| > 0.90$), where from any two highly correlated features, the feature with the highest mean correlation with the rest of the features was removed. Finally, the Boruta algorithm, a random forest feature selection method, was used to select important predictive features^[42,43]. The Boruta algorithm duplicated all features and shuffled the values in the so-called shadow features. Random forest classifiers were trained on the real and shadow features, and the algorithm subsequently compared the importance score of each feature and selected only those features where the importance of the real feature was higher compared with the shadow's feature importance [44]. Random forest classification models were trained on the 70% of the training data and tested on the remaining 30% of the training data. The best performing

radiomics models according to the summation of AUC and sensitivity value based on the test data in all strategies were selected and validated on the external validation data. All random forest parameters were set at default (Table S1) values. Figure 2 shows the radiomics workflow used in this study. Additionally, the range of the AUC values in the training data set is presented.

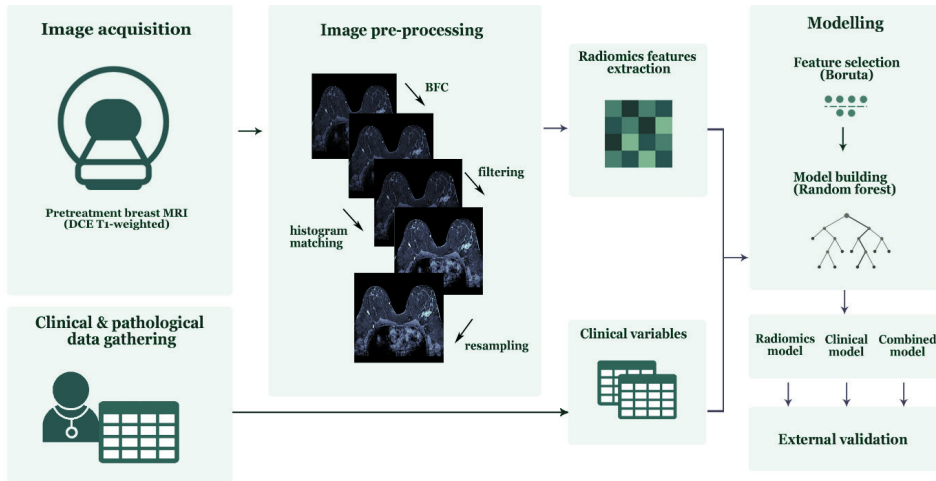


Figure 2:

Radiomics workflow used in this study. Abbreviations, MRI = Magnetic Resonance Imaging, DCE = Dynamic Con-trast-Enhanced, BFC = Bias Field Correction.

Clinical and Combined Model Development

Clinical and combined (based on radiomics features and clinical variables) random forest models were trained, tested, and validated using the same strategy used to develop the radiomics models as described above. Clinical models were based on the available clinical characteristics, including age, clinical tumor stage (cT), clinical nodal stage (cN), clinical tumor grade, tumor histology, and breast cancer subtype. The best performing clinical and combined models according to the summation of AUC and sensitivity value based on the test data in all strategies were selected and validated on the external validation data. All random forest parameters were set as default. Additionally, the range of the AUC values in the training data set was presented.

Statistical Analysis

Image pre-processing steps were performed in Python (version 3.7) using an in-house developed pipeline based on the computer vision packages opencv (version 4.1.0), SimpleITK (version 1.2.0), and numpy (version 1.16.2) procedure. The remaining statistical analysis, feature selection, model development, and model evaluation were performed in R (version 3.6.3) using R studio (version 1.2.1335, Vienna, Austria) ^[45]

and the R packages Boruta (version 7.0.0), Caret (version 6.0–85), Smotefamily (version 1.3.1), RandomForest (version 4.6–14), and pROC, (version 1.3.1) [46]. The difference between cohorts was assessed using independent samples *t*-test for continuous normally distributed variables, and Pearson chi-squared test for categorical variables. Statistical significance was based on *p*-values < 0.05 for both tests. The models developed were evaluated using the AUC and the 95% confidence interval (CI). DeLong's test was used to compare AUC values. In addition, the sensitivity and specificity and the negative predicted value (NPV) and positive predictive value (PPV) were derived from the confusion matrix. The radiomics quality score (RQS) was used to assess the radiomics workflow [14]. This study checked the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnoses (TRIPOD) guidelines [47,48].

Table 2: Clinical patient and tumor characteristics of patients in both complete data from the Maastricht University Medical Center+ (MUMC+) and Zuyderland Medical Center (ZMC) hospital.

Characteristics	MUMC+	ZMC	<i>p</i> -Value
Number of patients	129	161	-
Patient Age (years) (mean; range)	51 (28–73)	52 (28–79)	0.378
Number of tumors	152	168	-
Clinical tumor stage (%)			0.007
T1	29 (19.1)	16 (9.5)	
T2	99 (65.1)	103 (61.3)	
T3	20 (13.2)	37 (22.0)	
T4	4 (2.6)	12 (7.2)	
Clinical nodal stage (%)			<0.001
N0	88 (57.9)	59 (35.1)	
N1	44 (29.0)	87 (51.8)	
N2	9 (5.9)	12 (7.1)	
N3	11 (7.2)	7 (4.2)	
Unknown	0 (0.0)	3 (1.8)	
Clinical tumor grade (%)			0.003
1	8 (5.3)	22 (13.1)	
2	70 (46.1)	84 (50.0)	
3	68 (44.7)	62 (36.9)	
Unknown	6 (3.9)	0 (0.0)	
Tumor histology (%)			0.009
Invasive ductal carcinoma	136 (89.5)	134 (79.8)	
Invasive lobular carcinoma	10 (6.6)	14 (8.3)	
Invasive mixed ductal/lobular carcinoma	0 (0.0)	9 (5.4)	
Other invasive carcinoma	6 (3.9)	11 (6.5)	
Cancer Subtype (%)			0.921
HR+ and HER2-	80 (52.6)	82 (48.8)	
HR+ and HER2+	22 (14.5)	26 (15.5)	
HR- and HER2+	19 (12.5)	22 (13.1)	
Triple-negative	31 (20.4)	38 (22.6)	
Response to NAC (%)			0.331
pCR	53 (34.9)	49 (29.2)	
Non-pCR	99 (65.1)	119 (70.8)	

Abbreviations, HR = Hormone Receptor, HER2 = Human Epidermal growth factor Receptor 2.

Results

Patients Demographics

A total of 322 women with invasive breast cancer and treated with NST were considered for inclusion, of whom 32 were excluded (Figure 1B). A total of 290 women with 320 breast tumors met the inclusion criteria, of whom 129 women with 152 breast tumors were collected at the MUMC+ and 161 women with 168 breast tumors at the ZMC. Table 2 summarizes the patient and tumor characteristics of both datasets. The pCR rate of the included tumors was 34.9% (53/152) and 29.2% (49/168) in the MUMC+ and ZMC cohorts, respectively, showing no significant difference. There were significant cohort differences in clinical tumor stage, clinical nodal stage, clinical tumor grade, and tumor histology (Table 3). Clinical tumor stage, clinical tumor grade, and breast cancer subtype showed significant differences between pCR and non-pCR tumors within the individual cohorts (Table 3).

The results reported in the manuscript are based on the in-house developed image preprocessing pipeline, whereas the results based on the image pre-processing proposed by Pyradiomics are reported in the Supplementary Materials (Tables S2 and S3 and Figure S1). In both the radiomics and combined models, no significant differences were found (Table S4).

Table 3: Clinical patient and tumor characteristics of patients in both complete data cohorts on pCR and non-pCR tumors from the Maastricht University Medical Center (MUMC+) and Zuyderland Medical Center (ZMC) hospitals.

Characteristics	MUMC+			ZMC		
	Non-pCR	pCR	<i>p</i> -Value	Non-pCR	pCR	<i>p</i> -Value
Number of tumors	99	53	-	119	49	-
Patient Age (years) (mean; range)	52 (32–72)	51 (28–73)	0.600	53 (31–79)	52 (28–73)	0.538
Clinical tumor stage (%)			0.019*			0.023
T1	12 (12.1)	17 (32.1)		6 (5.0)	10 (20.4)	
T2	68 (68.7)	31 (58.5)		76 (63.9)	27 (55.1)	
T3	16 (16.2)	4 (7.5)		28 (23.5)	9 (18.4)	
T4	3 (3.0)	1 (1.9)		9 (7.6)	3 (6.1)	
Clinical nodal stage (%)			0.943			0.526
N0	56 (56.6)	32 (60.3)		39 (32.8)	20 (40.8)	
N1	29 (29.3)	15 (28.3)		62 (52.1)	25 (51.0)	
N2	6 (6.1)	3 (5.7)		11 (9.2)	1 (2.0)	
N3	8 (8.1)	3 (5.7)		5 (4.2)	2 (4.1)	
Unknown	0 (0.0)	0 (0.0)		2 (1.7)	1 (2.0)	
Clinical tumor grade (%)			<0.001*			0.002
1	8 (8.1)	0 (0.0)		19 (15.9)	3 (6.1)	
2	58 (58.6)	12 (22.7)		66 (55.5)	18 (36.7)	
3	32 (32.3)	36 (67.9)		34 (28.6)	28 (57.2)	
Unknown	1 (1.0)	5 (9.4)		0 (0.0)	0 (0.0)	

Table 3: continued.

Characteristics	MUMC+			ZMC		
	Non-pCR	pCR	p-Value	Non-pCR	pCR	p-Value
Tumor histology (%)			0.913			0.030
Invasive ductal carcinoma	89 (89.9)	47 (88.7)		91 (76.5)	43 (87.8)	
Invasive lobular carcinoma	6 (6.1)	4 (7.5)		13 (10.9)	1 (2.0)	
Invasive mixed ductal/lobular carcinoma	0 (0.0)	0 (0.0)		9 (7.6)	0 (0.0)	
Other invasive carcinoma	4 (4.0)	2 (3.8)		6 (5.0)	5 (10.2)	
Cancer Subtype (%)			<0.001*			<0.001
HR+ and HER2-	64 (64.6)	16 (30.2)		75 (63.0)	7 (14.3)	
HR+ and HER2+	15 (15.2)	7 (13.2)		14 (11.8)	12 (24.5)	
HR- and HER2+	6 (6.1)	13 (24.5)		5 (4.2)	17 (34.7)	
Triple-negative	14 (14.1)	17 (32.1)		25 (21.0)	13 (26.5)	

Abbreviations, pCR = pathologic Complete Response, HR = Hormone Receptor, HER2 = Human Epidermal growth factor Receptor 2.

Radiomics Models—Feature Selection and Model Performance

Of the 833 features extracted per ROI, 87 features were removed, as they were reported to be significantly affected by inter-observer segmentation variability (Table S5). In the best performing radiomics models in all strategies, one feature (*firstorder_maximum*) was removed, as it showed near zero variance. This was followed by the removal of: 574, 568, and 568 highly correlated features in strategy 1, 2, and 3, respectively, leaving 172, 178, and 178 features in the respective cohorts. The Boruta algorithm selected 5, 1, and 6 features in the best performing radiomics models for strategy 1, 2, and 3, respectively (Table 4A).

The results of the best performing radiomics models developed in the three strategies are shown in Table 5A. The AUC values in the validation cohorts were 0.55 (95% CI: 0.46–0.65), 0.52 (95%CI: 0.42–0.62), and 0.50 (95%CI: 0.37–0.64) for the respective strategies 1, 2, and 3. The sensitivity values ranged between 24% and 73% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.46 and 0.86 (Table S6).

Table 4: Selected features in best performing radiomics, clinical, and combined models for the three strategies.

	Strategy 1	Strategy 2	Strategy 3
A (Radiomics)	O_glszm_GrayLevelVariance	W.LHH_firstorder_Kurtosis	O_shape_Sphericity
	W.HLL_firstorder_Mean		W.LLH_glszm_GrayLevel-Non-Uniformity
	W.HLL_glcm_Imc1		W.LLH_glszm_ZoneEntropy
	W.HLH_glcm_InverseVariance		W.HHL_glcm_Imc1
	W.LLL_ngtdm_Complexity		W.HHH_glrIm_RunEntropy W.LLL_glcm_DifferenceVariance
B (Clinical)	Age	cT	Age
	cT	cN	cT
	Tumor grade	Tumor grade	Tumor grade
	Breast cancer subtype	Breast cancer subtype	Breast cancer subtype
C (Combined)	Tumor grade	Tumor grade	cT
	Breast cancer subtype	Breast cancer subtype	Tumor grade
	O_shape_Sphericity	W.LHL_firstorder_kurtosis	Breast cancer subtype
	O_firstorder_Mean	W.HHL_gldm_Dependence-Variance	O_shape_Sphericity
	W.HLL_glcm_Imc2		W.LLH_glszm_SmallAreaLowGrayLevelEmphasis
	W.HLL_glszm_ZoneEntropy		
	W.HLH_glcm_InverseVariance		

Abbreviations: O = original, W = wavelet, cT = clinical tumor stage, and cN = clinical nodal stage.

Clinical Models—Feature Selection and Model Performance

The clinical variables available were patient age, cT, cN, clinical tumor grade, tumor histology, and breast cancer subtype. None of the clinical variables were highly correlated. The Boruta algorithm selected four features in the best performing clinical models for all strategies (Table 4B). The results of the clinical models performed in the three settings are shown in Table 5B. The AUC values in the validation cohorts were 0.71 (95% CI: 0.62–0.79), 0.77 (95% CI: 0.70–0.85), and 0.72 (95% CI: 0.61–0.83) for strategy 1, 2, and 3, respectively. The clinical models performed significantly better compared with the radiomics models (Figure 3). The sensitivity values ranged between 41% and 47% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.68 and 0.88 (Table S6).

Combined Models—Feature Selection and Model Performance

Of the 833 features extracted per ROI, 87 features were removed, as they were reported to be significantly affected by inter-observer segmentation variability. In the best performing combined models in all strategies, one feature (*firstorder_maximum*) was removed, as it showed near zero variance. This was followed by the removal of 580, 563, and 577 highly correlated features in strategy 1, 2 and 3, respectively, leaving 172, 189, and 175 features in the respective cohorts. The Boruta algorithm selected 7, 4, and 6 features in the best performing radiomics models for strategy 1, 2, and 3, respectively (Table 4C). The three models all contained the same clinical features, clinical tumor grade, and clinical breast cancer subtype. The results of the best performing combined models developed in the three strategies are shown in Table 5C. The AUC values in the validation cohorts were 0.73 (95% CI: 0.65–0.81), 0.69 (95%CI: 0.61–0.78), and 0.71 (95%CI: 0.60–0.81) for the respective strategies 1, 2, and 3. The sensitivity values ranged between 38% and 51% in the validation cohorts. The 100 radiomics models developed in the three strategies resulted in a range of AUC values in the training cohorts between 0.59 and 0.91 (Table S6).

RQS and TRIPOD Results

This study scored a RQS score of 41.7% (15 out of 36 points) (Table S7). The score of the TRIPOD checklist was 73% (24 out of 33 applicable items) (Table S8).

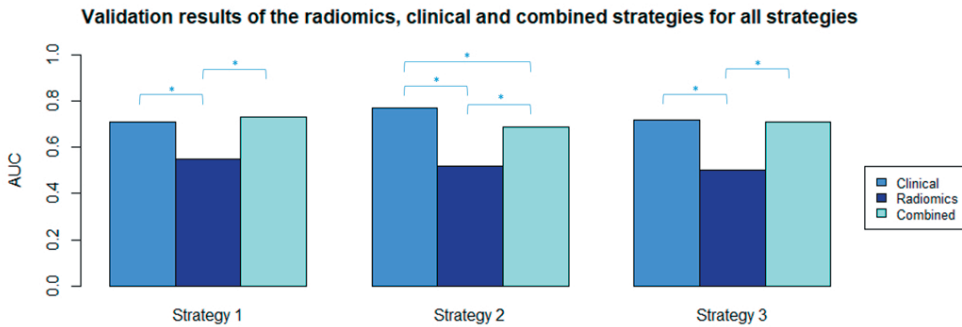


Figure 3:

AUC values from the selected radiomics, clinical, and combined validation models in all strategies. * Significant difference between AUC values with p -value < 0.05 (p -values were calculated using the ROC test by Delong method).

Table 5: Performance of best performing random forest radiomics (5A), clinical (5B), and combined (5C) models for the three strategies.

	Strategy 1				Strategy 2				Strategy 3					
	Training MUMC+		Validation ZMC		Training ZMC		Validation MUMC+		Training MUMC+		Validation 70% Mixed		Validation 30% Mixed	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
A (Radiomics)														
Area under the ROC	0.71	0.78	0.55	0.67	0.64	0.67	0.52	0.65	0.60	0.65	0.50	0.50	0.50	0.50
95% CI	0.59–0.82		0.63–0.92		0.54–0.75		0.42–0.62		0.49–0.71		0.51–0.80		0.37–0.64	
Sensitivity (%)	53	59	73	60	44	60	28	48	38	48	24	24	24	
Specificity (%)	89	79	36	72	75	72	62	77	92	77	88	88	88	
PPV (%)	70	63	32	47	42	47	28	48	69	48	47	47	47	
NPV (%)	79	76	77	81	77	81	62	77	75	77	72	72	72	
B (Clinical)														
Strategy 1														
Training MUMC+	Train		Test		Train		Test		Train		Test		Train	
Area under the ROC	0.79	0.81	0.71	0.84	0.81	0.84	0.77	0.86	0.75	0.86	0.72	0.72	0.72	0.72
95% CI	0.71–0.87		0.68–0.95		0.73–0.89		0.70–0.85		0.68–0.83		0.77–0.95		0.61–0.83	
Sensitivity (%)	54	86	45	71	54	71	47	71	52	71	41	41	41	
Specificity (%)	87	64	74	86	85	86	85	85	77	84	78	78	78	
PPV (%)	69	57	42	67	59	67	63	63	52	68	46	46	46	
NPV (%)	78	89	77	88	82	88	75	75	77	86	75	75	75	
Strategy 2														
Training MUMC+	Train		Test		Train		Test		Train		Test		Train	
Area under the ROC	0.82	0.83	0.73	0.86	0.79	0.86	0.69	0.86	0.79	0.86	0.71	0.71	0.71	
95% CI	0.74–0.90		0.70–0.97		0.71–0.88		0.61–0.78		0.73–0.86		0.76–0.96		0.60–0.81	
Sensitivity (%)	53	67	51	71	51	71	51	71	52	71	38	38	38	
Specificity (%)	88	88	82	82	87	82	67	67	85	89	83	83	83	
PPV (%)	69	77	53	63	62	63	45	45	61	75	50	50	50	
NPV (%)	78	82	80	88	81	88	72	72	79	87	75	75	75	
Strategy 3														
Training MUMC+	Train		Test		Train		Test		Train		Test		Train	
Area under the ROC	0.82	0.83	0.73	0.86	0.79	0.86	0.69	0.86	0.79	0.86	0.71	0.71	0.71	
95% CI	0.74–0.90		0.70–0.97		0.71–0.88		0.61–0.78		0.73–0.86		0.76–0.96		0.60–0.81	
Sensitivity (%)	53	67	51	71	51	71	51	71	52	71	38	38	38	
Specificity (%)	88	88	82	82	87	82	67	67	85	89	83	83	83	
PPV (%)	69	77	53	63	62	63	45	45	61	75	50	50	50	
NPV (%)	78	82	80	88	81	88	72	72	79	87	75	75	75	

Abbreviations, MUMC+ = Maastricht University Medical Center+, ZMC = Zuyderland Medical Center, CI = confidence interval, PPV = positive predicted value, NPV = negative predicted value.



Discussion

In this multicenter study, we investigated the value of pretreatment contrast-enhanced MRI-based radiomics for the prediction of pCR to NST in breast cancer patients using radiomics, clinical, and combined models in three different data-mixing strategies. The AUC values of the radiomics, clinical, and combined models in the validation datasets of the three strategies had ranges of 0.50–0.55, 0.71–0.77, and 0.69–0.73, respectively. Different radiomics features were selected for the radiomics and combined models in the three strategies, while the selected clinical features were mostly the same in all scenarios, with comparable performances. These results indicate poor performance of the radiomics features and that the radiomic features had no added value to the clinical models developed for the prediction of pCR to NST in breast cancer patients.

The clinical models significantly outperformed the radiomics models for the prediction of pCR to NST in all strategies. This indicates that radiomics features in these scenarios did not have an added value to the clinical model we developed. Furthermore, the variation in the features selected and model performance was greater in the radiomics models compared with the clinical models. However, based on current knowledge in the radiomics field, we cannot say that radiomics features do not have an added value unless the variations in acquisition and reconstruction parameters are properly addressed. Due to the lack of reproducibility data, this study could not analyze the effects of different acquisition and reconstruction parameters on radiomics feature values. Furthermore, the significant differences in population characteristics between the two cohorts could have led to the low performance of the radiomics models. While there was overlap in breast cancer phenotypes, the proportions at which these phenotypes occur may have differed so that the differences in prevalence resulted in differences in overall classification performances.

The results of this study indicate that even extensive MRI pre-processing and homogenization of the MR images do not sufficiently address the variations in acquisition and reconstruction parameters. This is in line with studies published in recent years that investigated the reproducibility of MRI radiomics features in test-retest phantom data as well as in patient data of varying disease sites, and showed that, among others, the variations in acquisition and reconstruction parameters strongly influence the values (concordance) of radiomics features [24,27–29,49–52]. Shur et al. [29] performed a test-retest 1.5T MRI phantom study using the same imaging protocol and showed that 20% of the examined features were not repeatable. A study on repeatability and reproducibility using a T2W pelvic phantom showed that radiomics features values are not only affected by varying acquisition parameters but also by the use of different MRI vendors and magnetic field strengths, wherein the reproducibility of the radiomic features is more

affected by difference in MRI vendor than by difference in magnetic field strength ^[49]. Overall, they reported that only 3.3% (31/944) of the examined features showed excellent robustness (ICC and CCC > 0.9). The radiomics community is currently trying to address these major hurdles.

Investigating comparable published work, we found a number of studies using only univariate predictive features without an external validation data cohort ^[18–21,53,54] and more recent published papers that were focusing on multivariate prediction models ^[32,33,55,56]. Hope Cain et al. [55] achieved an AUC value of 0.71 (95% CI: 0.58–0.83) for predicting pCR to NST in TN/HER2+ breast cancer patients; however, the model was not externally validated. Therefore, we anticipate that the results could not be generalized to scans acquired with different vendors/parameters than those used in the study. The study by Liu et al. [57] was the only study performing external radiomics model validation for the prediction of pCR to NST in breast cancer patients. The study differed from our research by the use of multiparametric (T2-weighted, diffusion-weighted images, and contrast-enhanced T1-weighted) MRI. However, the use of multiple MRI sequences for pCR prediction achieved better outcome with validation AUC values between 0.71 and 0.80. However, it is remarkable that their external validation results were obtained with MRI images that were much less extensively pre-processed compared to our images.

Our study also has its limitations. First, selection bias in retrospective studies is inevitable and so are the biases introduced by clinical protocols, such as HER2+ tumors receiving additional treatment compared to other tumors. Second, since the effect of different MRI scanners and acquisition and reconstruction parameters on radiomics features in breast imaging is not determined, we could not adjust our model for the potential variance induced by these factors in the radiomics feature values. Therefore, since data were collected from two hospitals using five MRI scanners with different acquisition and reconstruction parameters, noise may have been introduced into the models by incorporating radiomics features not robust to these variations. Third, while we believe that MRI preprocessing is a necessary step toward comparable images with intensity values having similar tissue meaning, it is possible that with our choice of preprocessing steps, consistent with current literature, we may have inadvertently removed quantitative information. However, the results obtained with the widely used pre-processing method proposed by Pyradiomics showed no significant differences from the result reported here. Fourth, the number of patients included in this study did not allow us to perform a subanalysis for the different breast cancer subtypes. Fifth, the data were collected over a relatively long period of time during which optimization of MRI acquisitions protocols occurred, which may have introduced variations as well. Last, for these analyses it was specifically chosen to use the peak-enhanced (2 min) post-contrast T1W images, as breast tumors are most visible on them and because some of the tumors

included cannot be seen on other sequences; for example, mucinous tumors and some of the invasive lobular tumors are not or only weakly visible on the subtraction images. In our opinion, performing the analysis using the subtraction images instead of the peak-enhanced images would have resulted in a significant decrement in the number of patients that could be analyzed. Furthermore, as the effects of the different breast MRI sequences on the radiomics features is not yet understood, future radiomics research in the field of breast cancer could focus on the use of the different MRI sequences, as well as on multiparametric and delta radiomics approaches.

Conclusions

In conclusion, this study showed no contribution of pretreatment contrast-enhanced MRI-based radiomics for the prediction of tumor pCR on NST in breast cancer patients, as neither the radiomics nor the combined models performed significantly better than the clinical models. However, without analysis of the effects of variations in acquisition and reconstruction parameters, it is currently not possible to conclude that pretreatment contrast-enhanced MRI-based radiomic features have no value in the prediction of pCR to NST. The effects of different acquisition and reconstruction parameters on radiomics feature values in breast imaging should be explored in future MRI-breast reproducibility studies to investigate whether further research into pretreatment MRI-based radiomics for the prediction of pCR to NST in breast cancer patients is useful.

Supplementary Materials

The following are available online at www.mdpi.com/xxx/s1, Supplementary Materials include Table S1: Default random forest parameters, Table S2: Selected features using the proposed image pre-processing method by Pyradiomics, Table S3: Results of the best performing models using the proposed image pre-processing method by Pyradiomics, Figure S1: Comparisons of the validation AUC values using the proposed image pre-processing method by Pyradiomics, Table S4: Comparison of the validation results of both pre-processing methods, Table S5: List of excluded features affected by inter-observer segmentation variability, Table S6: Ranked AUC values of the 100 radiomics, clinical, and combined trainings models for the three strategies, Table S7: Radiomics Quality Score, and Table S8: TRIPOD Checklist.

Author Contributions

Conceptualization, R.W.Y.G.; methodology, R.W.Y.G. and A.I.; software, S.P.P.; validation, R.W.Y.G., and A.I.; formal analysis, R.W.Y.G., A.I., and H.C.W.; investigation, R.W.Y.G.; resources, R.W.Y.G., S.S. and T.J.A.v.N.; data curation, R.W.Y.G., T.J.A.v.N., and M.B.I.L.; writing—original draft preparation, R.W.Y.G. and A.I.; writing—review and editing, all authors; visualization, R.W.Y.G. and S.P.P.;

supervision, M.L.S. and H.C.W.; project administration, R.W.Y.G.; funding acquisition, M.L.S. All authors have read and agreed to the published version of the manuscript.

Funding

Renée W.Y. received a salary from Kankeronderzoekfonds Limburg. Authors acknowledge financial support from an ERC advanced grant (ERC-ADG-2015 n° 694812–Hypoximmuno). Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172 and EuCanImage n° 952103.

Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki and was approved by the Institutional Review Board (or Ethics Committee) of the University Hospital Maastricht and Maastricht University (METC azM/UM) (Ethic code: 2017-0152 with date of approval 12 October 2017).

Informed Consent Statement

Patient consent was waived due to the retrospective design of the study.

Data Availability Statement

The data presented in this study are available on reasonable request from the corresponding author. Due to privacy restrictions the data are not publicly available.

Conflicts of Interest

P.L. reports, within the submitted work, grants/sponsored research agreements from radiomics SA, ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in-kind manpower contribution from radiomics SA, BHV, Varian, Elekta, ptTheragnostic, BMS. Dr Lambin has minority shares in the company radiomics SA, MedC2; he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA; three non-patented invention (software) licensed to ptTheragnostic/DNAmito, Radiomics SA, and Health Innovation Ventures; and non-licensed patents on a Lymphocytes Sparing radiotherapy (P126537PC00), Radiomics hypoxia: P125078US00, Image preprocessing for AI (P124711PC00). He confirms that none of the above entities or funding was involved in the preparation of this paper. The other authors declare no conflicts of interest.

References

1. Vugts, G.; Maaskant-Braat, A.J.; Nieuwenhuijzen, G.A.; Roumen, R.M.; Luiten, E.J.; Voogd, A.C. Patterns of Care in the Administration of Neo-adjuvant Chemotherapy for Breast Cancer. A Population-Based Study. *Breast J.* **2016**, *22*, 316–321.
2. Murphy, B.L.; Day, C.N.; Hoskin, T.L.; Habermann, E.B.; Boughey, J.C. Neoadjuvant Chemotherapy Use in Breast Cancer is Greatest in Excellent Responders: Triple-Negative and HER2+ Subtypes. *Ann. Surg. Oncol.* **2018**, *25*, 2241–2248.
3. Spronk, P.E.R.; de Ligt, K.M.; van Bommel, A.C.M.; et al. Current decisions on neoadjuvant chemotherapy for early breast cancer: Experts' experiences in the Netherlands. *Patient Educ. Couns.* **2018**, *101*, 2111–2115.
4. Loibl, S.; Denkert, C.; von Minckwitz, G. Neoadjuvant treatment of breast cancer—Clinical and research perspective. *Breast* **2015**, *24*, S73–S77.
5. Prevos, R.; Smidt, M.L.; Tjan-Heijnen, V.C.; et al. Pre-treatment differences and early response monitoring of neoadjuvant chemotherapy in breast cancer patients using magnetic resonance imaging: A systematic review. *Eur. Radiol.* **2012**, *22*, 2607–2616.
6. Lobbes, M.B.; Prevos, R.; Smidt, M.; et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: A systematic review. *Insights Imaging* **2013**, *4*, 163–175.
7. Weber, J.J.; Jochelson, M.S.; Eaton, A.; et al. MRI and Prediction of Pathologic Complete Response in the Breast and Axilla after Neoadjuvant Chemotherapy for Breast Cancer: MRI and Pathologic Complete Response. *J. Am. Coll. Surg.* **2017**, *225*, 740–746.
8. Bouzon, A.; Acea, B.; Soler, R.; et al. Diagnostic accuracy of MRI to evaluate tumour response and residual tumour size after neoadjuvant chemotherapy in breast cancer patients. *Radiol. Oncol.* **2016**, *50*, 73–79.
9. van der Noordaa, M.E.M.; van Duijnhoven, F.H.; Loo, C.E.; et al. Identifying pathologic complete response of the breast after neoadjuvant systemic therapy with ultrasound guided biopsy to eventually omit surgery: Study design and feasibility of the MICRA trial (Minimally Invasive Complete Response Assessment). *Breast* **2018**, *40*, 76–81.
10. Heil, J.; Sinn, P.; Richter, H.; et al. RESPONDER—Diagnosis of pathological complete response by vacuum-assisted biopsy after neoadjuvant chemotherapy in breast Cancer—A multicenter, confirmative, one-armed, intra-individually-controlled, open, diagnostic trial. *BMC Cancer* **2018**, *18*, 851.
11. van Loevezijn, A.A.; van der Noordaa, M.E.M.; van Werkhoven, E.D.; et al. Minimally Invasive Complete Response Assessment of the Breast After Neoadjuvant Systemic Therapy for Early Breast Cancer (MICRA trial): Interim Analysis of a Multicenter Observational Cohort Study. *Ann. Surg. Oncol.* **2020**, doi:10.1245/s10434-020-09273-0.
12. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107.
13. Ibrahim, A.; Primakov, S.; Beuque, M.; et al. Radiomics for precision medicine: Current challenges,

- future prospects, and the proposal of a new framework. *Methods* **2021**, *188*, 20–29.
14. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762.
 15. Larue, R.T.; Defraene, G.; De Ruyscher, D.; Lambin, P.; van Elmpt, W. Quantitative radiomics studies for tissue characterization: A review of technology and methodological procedures. *Br. J. Radiol.* **2017**, *90*, 20160665.
 16. Refaee, T.; Wu, G.; Ibrahim, A.; et al. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration* **2020**, *99*, 99–107.
 17. Ahmed, A.; Gibbs, P.; Pickles, M.; Turnbull, L. Texture analysis in assessment and prediction of chemotherapy response in breast cancer. *J. Magn. Reson. Imaging* **2013**, *38*, 89–101.
 18. Parikh, J.; Selmi, M.; Charles-Edwards, G.; et al. Changes in primary breast cancer heterogeneity may augment midtreatment MR imaging assessment of response to neoadjuvant chemotherapy. *Radiology* **2014**, *272*, 100–112.
 19. Teruel, J.R.; Heldahl, M.G.; Goa, P.E.; et al. Dynamic contrast-enhanced MRI texture analysis for pretreatment prediction of clinical and pathological response to neoadjuvant chemotherapy in patients with locally advanced breast cancer. *NMR Biomed.* **2014**, *27*, 887–896.
 20. Chamming's F.; Ueno, Y.; Ferre, R.; et al. Features from Computerized Texture Analysis of Breast Cancers at Pretreatment MR Imaging Are Associated with Response to Neoadjuvant Chemotherapy. *Radiology* **2018**, *286*, 412–420.
 21. Yoon, H.J.; Kim, Y.; Chung, J.; Kim, B.S. Predicting neo-adjuvant chemotherapy response and progression-free survival of locally advanced breast cancer using textural features of intratumoral heterogeneity on F-18 FDG PET/CT and diffusion-weighted MR imaging. *The Breast J.* **2019**, *25*, 373–380.
 22. Baessler, B.; Weiss, K.; Pinto Dos Santos, D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest. Radiol.* **2019**, *54*, 221–228.
 23. Mackin, D.; Fave, X.; Zhang, L.; et al. Measuring CT scanner variability of radiomics features. *Investig. Radiol.* **2015**, *50*, 757.
 24. Rai, R.; Holloway, L.C.; Brink, C.; et al. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Med. Phys.* **2020**, *47*, 3054–3063.
 25. Ibrahim, A.; Refaee, T.; Primakov, S.; et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848.
 26. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS ONE* **2021**, *16*, e0251147.
 27. Dreher, C.; Kuder, T.A.; Konig, F.; et al. Radiomics in diffusion data: A test-retest, inter- and intra-reader DWI phantom study. *Clin. Radiol.* **2020**, *75*, 798.e13–798.e22.
 28. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci. Rep.* **2019**, *9*, 4800.
 29. Shur, J.; Blackledge, M.; D Arcy, J.; et al. MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study. *Eur. Radiol. Exp.* **2021**, *5*, 1–11.

Chapter 3

30. Ogston, K.N.; Miller, I.D.; Payne, S.; et al. A new histological grading system to assess response of breast cancers to primary chemotherapy: Prognostic significance and survival. *Breast* **2003**, *12*, 320–327.
31. Pinder, S.E.; Provenzano, E.; Earl, H.; Ellis, I.O. Laboratory handling and histology reporting of breast specimens from patients who have received neoadjuvant chemotherapy. *Histopathology* **2007**, *50*, 409–417.
32. Fan, M.; Wu, G.; Cheng, H.; Zhang, J.; Shao, G.; Li, L. Radiomic analysis of DCE-MRI for prediction of response to neoadjuvant chemotherapy in breast cancer patients. *Eur. J. Radiol.* **2017**, *94*, 140–147.
33. Braman, N.; Etesami, M.; Prasanna, P.; et al. Intratumoral and peritumoral radiomics for the pretreatment prediction of pathological complete response to neoadjuvant chemotherapy based on breast DCE-MRI. *Breast Cancer Res.* **2017**, *19*, no pagination, doi:10.1186/s13058-017-0846-1.
34. Granzier, R.W.Y.; Verbakel, N.M.H.; Ibrahim, A.; et al. MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Sci. Rep.* **2020**, *10*, 14163.
35. Nyul, L.G.; Udupa, J.K.; Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* **2000**, *19*, 143–150.
36. Hoebel, K.V.; Patel, J.B.; Beers, A.L.; et al. Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiol. Artif. Intell.* **2021**, *3*, doi:10.1148/ryai.2020190199.
37. Ligeró, M.; Jordi-Ollero, O.; Bernatowicz, K.; et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur. Radiol.* **2021**, *31*, 1460–1470.
38. Moradmand, H.; Aghamiri, S.M.R.; Ghaderi, R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J. Appl. Clin. Med. Phys.* **2020**, *21*, 179–190.
39. Sun, X.; Shi, L.; Luo, Y.; et al. Histogram-based normalization technique on human brain magnetic resonance images from different acquisitions. *Biomed. Eng. Online* **2015**, *14*, 73.
40. Senthilkumar, N.; Thimmiraja, J. Histogram Equalization for Image Enhancement Using MRI Brain Images. In Proceedings of the 2014 World Congress on Computing and Communication Technologies, Trichirappalli, India, 27 February–1 March 2014; pp. 80–83.
41. Duron, L.; Balvay, D.; Vande Perre, S.; et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS ONE* **2019**, *14*, e0213459.
42. Kursá, M.B.; Rudnicki, W.R. Feature Selection with the Boruta Package. *J. Stat. Softw.* **2010**, *36*, 1–13.
43. Kursá, M.B.; Jankowski, A.; Rudnicki, W.R. Boruta—A System for Feature Selection. *Fund. Inform.* **2010**, *101*, 271–286.
44. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958.
45. Racine, J.S. RStudio: A platform-independent IDE for R and Sweave. *J. Appl. Econom.* **2012**, *27*, 167–172.
46. Robin, X.; Turck, N.; Hainard, A.; et al. pROC: An open-source package for R and S+ to analyze and

- compare ROC curves. *BMC Bioinform.* **2011**, *12*, 77.
47. Moons, K.G.; Altman, D.G.; Reitsma, J.B.; et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73.
 48. Collins, G.S.; Reitsma, J.B.; Altman, D.G.; Moons, K.G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *BMC Med.* **2015**, *13*, 1.
 49. Bianchini, L.; Botta, F.; Origgi, D.; et al. PETER PHAN: An MRI phantom for the optimisation of radiomic studies of the female pelvis. *Phys. Med.* **2020**, *71*, 71–81.
 50. Schwier, M.; van Griethuysen, J.; Vangel, M.G.; et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci. Rep.* **2019**, *9*, 9441.
 51. Fiset, S.; Welch, M.L.; Weiss, J.; et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiother. Oncol.* **2019**, *135*, 107–114.
 52. Scalco, E.; Belfatto, A.; Mastropietro, A.; et al. T2w-MRI signal normalization affects radiomics features reproducibility. *Med. Phys.* **2020**, *47*, 1680–1691.
 53. Choudhery, S.; Gomez-Cardona, D.; Favazza, C.P.; et al. MRI Radiomics for Assessment of Molecular Subtype, Pathological Complete Response, and Residual Cancer Burden in Breast Cancer Patients Treated With Neoadjuvant Chemotherapy. *Acad. Radiol.* **2020**, in press.
 54. Henderson, S.; Purdie, C.; Michie, C.; et al. Interim heterogeneity changes measured using entropy texture features on T2-weighted MRI at 3.0 T are associated with pathological response to neoadjuvant chemotherapy in primary breast cancer. *Eur. Radiol.* **2017**, *27*, 4602–4611.
 55. Cain, E.H.; Saha, A.; Harowicz, M.R.; Marks, J.R.; Marcom, P.K.; Mazurowski, M.A. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: A study using an independent validation set. *Breast Cancer Res. Treat.* **2019**, *173*, 455–463.
 56. Xiong, Q.; Zhou, X.; Liu, Z.; et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin. Transl. Oncol.* **2020**, *22*, 50–59.
 57. Liu, Z.; Li, Z.; Qu, J.; et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: A multicenter study. *Clin. Cancer Res.* **2019**, *25*, 3538–3547.

A large, abstract blue watercolor splash with a white number 4 in the center. The splash is composed of various shades of blue, from light to dark, with some darker spots and a textured, organic appearance. The number 4 is a simple, bold, white sans-serif font, centered within the splash.

4

Chapter 4

Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer

Authors

Sanaz Samiei, Renée W. Y. Granzier, Abdalla Ibrahim, Sergey Primakov, Marc B. I. Lobbes, Regina G. H. Beets-Tan, Thiemo J. A. van Nijnatten, Sanne M. E. Engelen, Henry C. Woodruff and Marjolein L. Smidt

Adapted from

Cancers. 2021 Jan;13(4):757

DOI

10.3390/cancers13040757

Abstract

Radiomics features may contribute to increased diagnostic performance of MRI in the prediction of axillary lymph node metastasis. The objective of the study was to predict preoperative axillary lymph node metastasis in breast cancer using clinical models and radiomics models based on T2-weighted (T2W) dedicated axillary MRI features with node-by-node analysis. From August 2012 until October 2014, all women who had undergone dedicated axillary 3.0T T2W MRI, followed by axillary surgery, were retrospectively identified, and available clinical data were collected. All axillary lymph nodes were manually delineated on the T2W MR images, and quantitative radiomics features were extracted from the delineated regions. Data were partitioned patient-wise to train 100 models using different splits for the training and validation cohorts to account for multiple lymph nodes per patient and class imbalance. Features were selected in the training cohorts using recursive feature elimination with repeated 5-fold cross-validation, followed by the development of random forest models. The performance of the models was assessed using the area under the curve (AUC). A total of 75 women (median age, 61 years; interquartile range, 51–68 years) with 511 axillary lymph nodes were included. On final pathology, 36 (7%) of the lymph nodes had metastasis. A total of 105 original radiomics features were extracted from the T2W MR images. Each cohort split resulted in a different number of lymph nodes in the training cohorts and a different set of selected features. Performance of the 100 clinical and radiomics models showed a wide range of AUC values between 0.41–0.74 and 0.48–0.89 in the training cohorts, respectively, and between 0.30–0.98 and 0.37–0.99 in the validation cohorts, respectively. With these results, it was not possible to obtain a final prediction model. Clinical characteristics and dedicated axillary MRI-based radiomics with node-by-node analysis did not contribute to the prediction of axillary lymph node metastasis in breast cancer based on data where variations in acquisition and reconstruction parameters were not addressed.

Keywords

dedicated axillary MRI; axillary lymph node metastasis; node-by-node matching; radiomics; predictive modeling

Introduction

In breast cancer patients, the axillary lymph node status provides essential prognostic information about the locoregional recurrence and overall survival rate ^[1-4]. The five-year survival rate decreases from 99% to 85% with the presence of lymph node metastasis in the axilla ^[5]. The presence of axillary lymph node metastasis determines the extent of the surgical treatment plan, the potential need for (neo)adjuvant systemic therapy, and the possible indication for postmastectomy radiation therapy with regard to immediate breast reconstruction ^[6,7].

In the preoperative setting, imaging for axillary lymph node assessment is recommended in the clinical workup of invasive breast cancer patients ^[6]. For the evaluation of tumor extent in the breast or following neoadjuvant treatment, breast magnetic resonance imaging (MRI) is often performed, which includes the axilla in the field of view ^[8]. However, when using dedicated breast coils, the field of view of the axillary region can be limited ^[9]. Therefore, dedicated MR coils for visualization and assessment of the axillary region have been investigated ^[10-12]. Dedicated unenhanced T2-weighted (T2W) axillary MRI showed good diagnostic performance based on node-by-node analysis but remained insufficient to accurately exclude axillary lymph node metastasis ^[12].

Although preoperative imaging may be performed to guide the axillary management of patients, no current imaging modality with optimal diagnostic performance can replace the surgical axillary staging procedure. In the era of artificial intelligence, current developments in radiology focus on the improvement of decision support systems to maximize the potential role of noninvasive imaging modalities. Radiomics, the application of machine learning to medical imaging, is a rapidly evolving field that enables high-throughput quantitative data extraction from standard medical images in an automated fashion and subsequent data analysis, possibly combined with patient and tumor characteristics, improving the accuracy of diagnostic, predictive, and prognostic models ^[13,14]. The evaluation of the usefulness of radiomics based on mammography, ultrasound, and breast MRI has been explored, showing potential in axillary lymph node metastasis prediction ^[15-19]. However, this research focused on the prediction of axillary lymph node metastasis from the delineated breast tumor as the region of interest (ROI), and not from the lymph nodes themselves.

Accurate preoperative prediction of axillary lymph node metastasis in breast cancer patients can assist in clinical decision-making regarding the type of treatment. Radiomics features extracted from axillary lymph nodes may contribute to increased diagnostic performance of MRI in the prediction of metastasis. To our knowledge, no previous study has reported on node-by-node matching of axillary lymph nodes with pathological

findings in breast cancer patients in the field of radiomics. The purpose of this study was to predict preoperative axillary lymph node metastasis in breast cancer patients using clinical models and radiomics models based on unenhanced T2W dedicated axillary MRI features with node-by-node analysis.

Results

Patients Characteristics

A total of ninety women were considered for inclusion, of whom twelve were excluded due to treatment with neoadjuvant systemic therapy before axillary surgery and three with ductal carcinoma in situ only. Seventy-five patients (median age, 61 years; interquartile range, 51–68 years) with 511 axillary lymph nodes were included. Patient, tumor, and treatment characteristics are summarized in Table 1. The median number of axillary lymph nodes per patient was six, with a range of 1–18. Fourteen of the included patients were node-positive at final pathology, with a total of 36 axillary lymph nodes with macrometastases and 58 axillary lymph nodes without metastasis. The remaining 61 patients had 417 axillary lymph nodes without metastasis. The median number of voxels per ROI for all delineated axillary lymph nodes was 100 (interquartile range, 44–236) and 310 (interquartile range, 130–1676) for all delineated axillary lymph nodes with metastasis. The Spearman correlation between the number of voxels per ROI and the corresponding pathological outcome was 0.22.

Table I: Patient, tumor, and treatment characteristics.

Characteristic	Value
No. of patients	75
Age (years) (median; IQR)	61 (51–68)
Clinical tumor size (mm) (median, IQR)	19 (13–28)
Clinical tumor stage (%)	
T1	41 (54.7)
T2	32 (42.7)
T3	2 (2.6)
Clinical nodal stage (%)	
N0	68 (90.7)
N1	7 (9.3)
Tumor histology (%)	
Invasive ductal	55 (73.3)
Invasive lobular	11 (14.7)
Mixed invasive ductal & lobular	3 (4.0)
Other	6 (8.0)
Tumor grade (%)	
1	17 (22.7)
2	42 (56.0)
3	16 (21.3)

Table I: Patient, tumor, and treatment characteristics.

Characteristic	Value
Breast cancer subtype (%)	
ER + HER2-	55 (73.3)
ER + HER2+	6 (9.0)
ER - HER2+	2 (2.7)
Triple-negative	11 (14.7)
Not determined	1 (1.3)
Axillary surgery (%)	
SLNB	8 (10.7)
ALND	67 (89.3)

Abbreviations: ER, Estrogen receptor; HER2, Human epidermal growth factor receptor 2; IQR, interquartile range; SLNB, Sentinel lymph node biopsy; ALND, Axillary lymph node dissection.

Radiomics Feature Extraction and Model Development

A total of 105 original radiomics features were extracted from the dedicated axillary T2W MR images. No near-zero variance features were detected. Pearson pairwise correlation removed 53 highly correlated features. The optimal subset of features was selected in the training cohort using recursive feature elimination with repeated 5-fold cross-validation with a maximum of 20 features. Figure 1 shows the distribution of the number of selected features from the 100 iterations for the two different strategies (lymph nodes from all patients versus only lymph nodes from node-positive patients as data points) for each model. Supplementary Material A includes a list of how often each feature was chosen in the 100 iterations for each model.

As each iteration resulted in a different set of selected features for each model in both strategies, it was not possible to obtain a final prediction model. The minimum and maximum area under the curve (AUC) values in the training cohorts were 0.59–0.80, 0.60–0.85, 0.48–0.84, and 0.55–0.89 for models 1a, 1b, 2a, and 2b, respectively. The median AUC values for all models in the training cohorts were between 0.72–0.73. All models showed a wider range of AUC values in the validation cohorts. The AUC value distribution for all models in the training and validation cohorts are presented in the violin plots in Figure 2. The minimum and maximum sensitivity in the training cohorts were 30–66%, 53–83%, 7–74%, and 48–82% for models 1a, 1b, 2a, and 2b, respectively. The median sensitivity for all models in the training cohorts was between 47–66%. All models showed lower median sensitivity in the validation cohorts. The minimum and maximum PPV in the training cohorts were 46–78%, 55–83%, 25–80%, and 52–90% for models 1a, 1b, 2a, and 2b, respectively. The median PPV for all models in the training cohorts were between 61–67%. All models showed a lower median PPV in the validation cohorts. The diagnostic performance parameters of the radiomics models (100 iterations) are shown in Table 2.

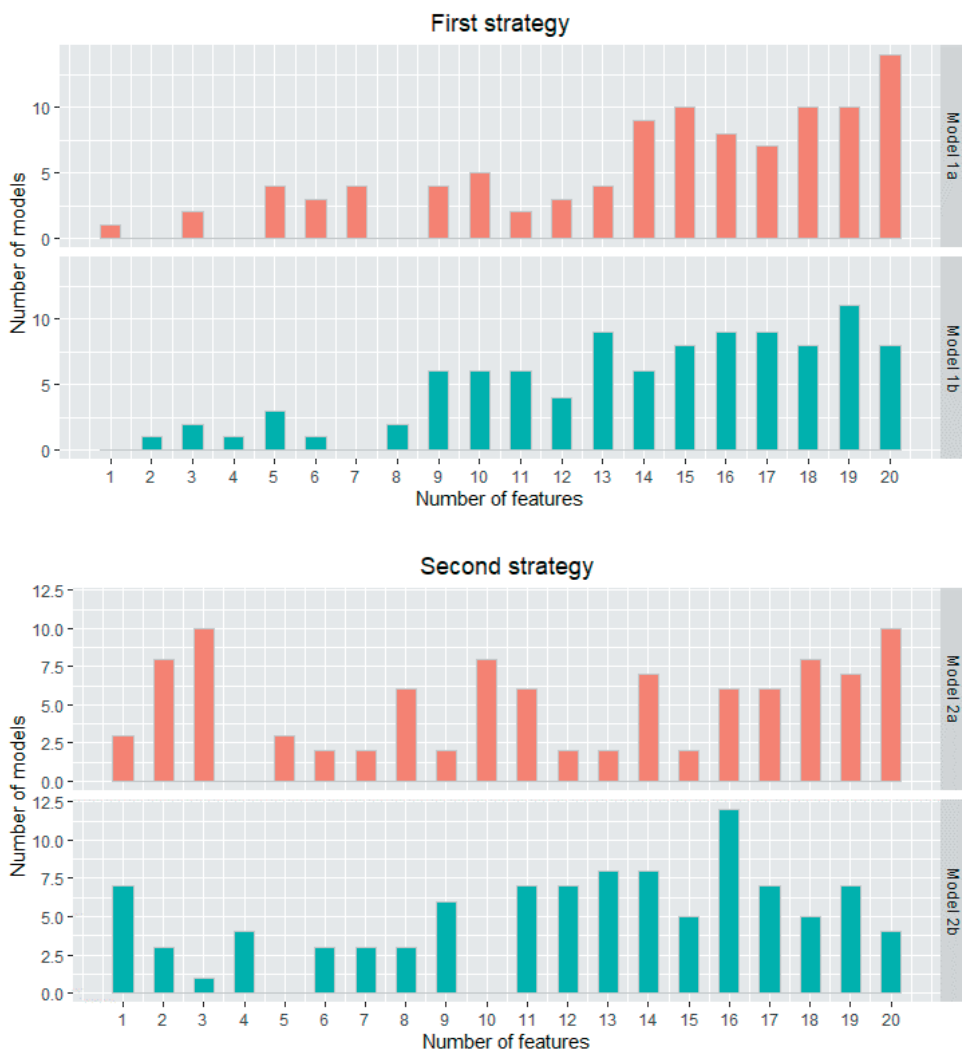
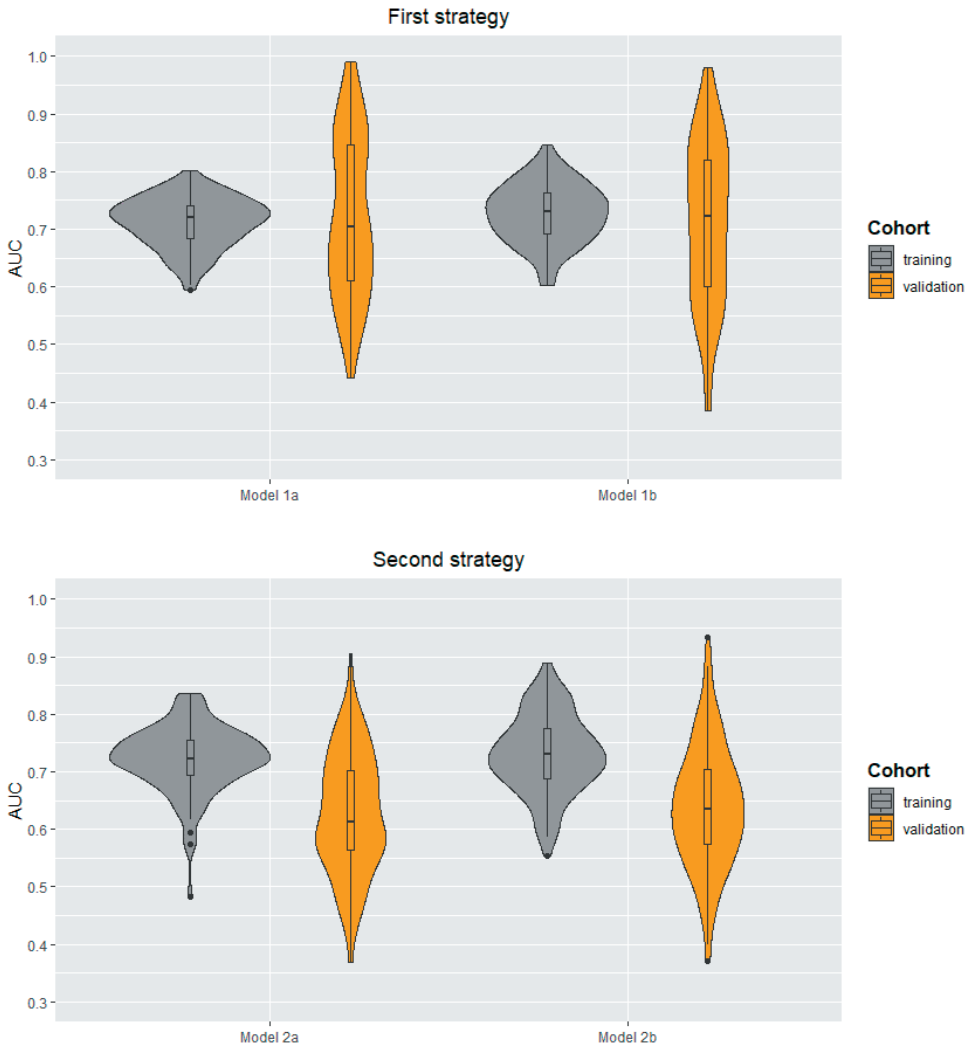


Figure I: First (A) and second (B) strategy: distribution of the number of features in each developed model. The two different models in both strategies were all developed 100 times.



4

Figure 2: Violin plots for the radiomics models developed using the first (A) and second (B) strategy: AUC value distribution (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort.

Table 2: The diagnostic performance of the radiomics models (100 iterations) for the first and second strategy.

Diagnostic parameters	Training			Validation			Training			Validation						
	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)				
First Strategy																
Model 1a						Model 1b										
Minimum	30	71	46	62	0	78	0	98	53	50	55	72	0	57	0	98
Median	47	81	61	72	33	90	2	99	66	67	67	80	50	75	1	99
Maximum	66	91	78	79	100	97	22	100	83	85	83	88	100	88	10	100
Second Strategy																
Model 2a						Model 2b										
Minimum	7	58	25	54	0	33	0	22	48	46	52	68	0	0	0	0
Median	50	81	62	74	33	76	50	71	66	68	67	80	64	60	50	75
Maximum	74	93	80	83	82	100	100	88	82	92	90	89	100	100	100	100

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; sens, sensitivity; spec, specificity.

The additional feature selection step with the cut-off values >0.75 , >0.80 , and >0.90 resulted in 44, 35, and 8 original features, respectively, available for recursive feature elimination with repeated 5-fold cross-validation. These results showed no differences compared to the results found without this additional feature selection step. The violin plots of the models developed after adding the additional feature selection step can be found in Figures S1–S3.

Radiomics Subanalysis

After the exclusion of ROIs with less than 50 voxels, a total of 71 patients were included for analyses, with 371 axillary lymph nodes. Thirteen of these patients were node-positive, with a total of 31 axillary lymph nodes with metastasis and 34 axillary lymph nodes without metastases. The remaining 58 patients had 340 axillary lymph nodes without metastasis. Excluding small lymph nodes resulted in balanced training cohorts in models 1a and 2a, eliminating the need to perform random undersampling (models 1b and 2b). The minimum and maximum AUC values of the balanced models 1a and 2a in the training and validation cohorts of this subanalysis were 0.53–0.82 and 0.41–0.83, respectively. Violin plots with the distribution of the AUC values and the diagnostic performance parameters of the subanalysis are provided in Table S1 and Figure S4.

Clinical Model Development

The following clinical characteristics were available and selected for the development of the clinical models: patient age, clinical tumor size, clinical tumor stage, tumor histology, tumor grade, and receptor subtype (ER, PR, and HER2+). No highly correlated clinical characteristics were present. The minimum and maximum AUC values in the training cohorts were 0.52–0.66, 0.43–0.71, 0.41–0.67, and 0.43–0.74 for models 1a, 1b, 2a, and 2b, respectively. The median AUC values for all models in the training cohorts were between 0.59–0.60. All models showed a wider range of AUC values in the validation cohorts. The AUC value distribution for all models in the training and validation cohorts are presented in the violin plots in Figure 3. The minimum and maximum sensitivity in the training cohorts were 18–64%, 31–71%, 0–65%, and 33–73% for models 1a, 1b, 2a, and 2b, respectively. The median sensitivity for all models in the training cohorts was between 42–58%. All models showed lower median sensitivity in the validation cohorts, except for model 2b. The minimum and maximum positive predictive value (PPV) in the training cohorts were 42–71%, 41–85%, 48–73%, and 43–86% for models 1a, 1b, 2a, and 2b, respectively. The median PPV for all models in the training cohorts was between 68–70%. All models showed a lower median PPV in the validation cohorts, except for model 2a. In all four models, the clinical tumor size was ranked as the most important clinical characteristic followed by age. The diagnostic performance parameters of the clinical models (100 iterations) are shown in Table 3.

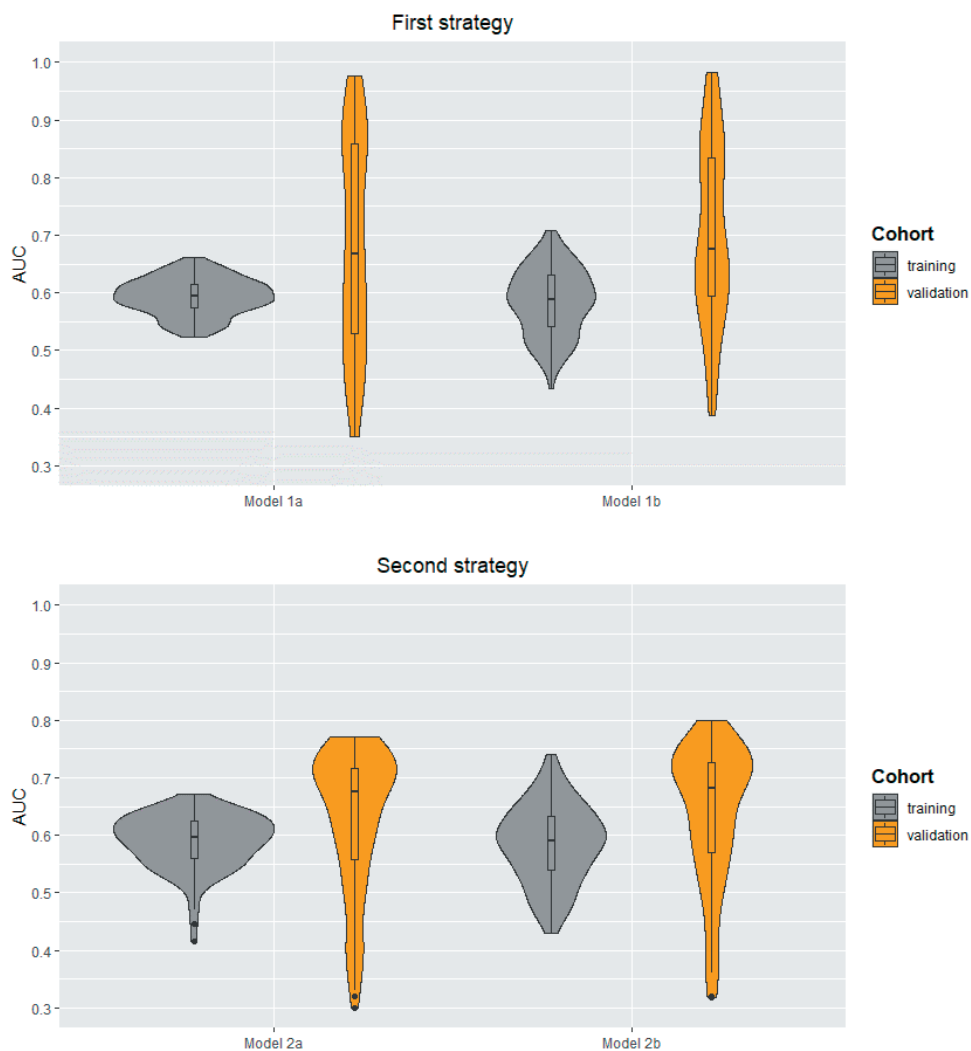


Figure 3: Violin plots for the clinical models developed using the first (A) and second (B) strategy: AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort.

Table 3: The diagnostic performance of the clinical models (100 iterations) for the first and second strate-gy.

Diagnostic parameters	Training			Validation			Training			Validation						
	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)	Sens (%)	Spec (%)	PPV (%)	NPV (%)				
First Strategy																
Model 1a						Model 1b										
Minimum	18	64	42	65	0	40	0	99	31	46	41	42	0	14	0	97
Median	50	86	68	72	0	91	0	99	58	74	70	64	50	64	1	99
Maximum	64	93	71	78	100	99	18	100	71	92	85	73	100	88	9	100
Second Strategy																
Model 2a						Model 2b										
Minimum	0	55	48	61	0	0	10	34	33	45	43	43	0	0	10	0
Median	42	85	68	72	39	80	69	73	57	75	70	63	61	53	43	67
Maximum	65	100	73	80	100	100	73	84	73	91	86	74	100	100	100	86

Abbreviations: NPV, negative predictive value; PPV, positive predictive value; sens, sensitivity; spec, specificity.

RQS and TRIPOD

This study scored a radiomics quality score (RQS) of 58% (21 out of 36 points) (Table S2). The score of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist was 67% (18 out of 27 applicable items) (Table S3).

Discussion

Accurate preoperative prediction of axillary lymph node metastasis can assist in clinical decision-making regarding the extent of axillary surgery and radiation therapy, and provide essential prognostic information. In this study, clinical models and radiomics models based on T2-weighted dedicated axillary MRI features with node-by-node analysis were investigated for the preoperative prediction of axillary lymph node metastasis. The different sets of features selected at each split resulted in a wide range of AUC values and did not allow for the development of a final radiomics prediction model. The performance of the clinical models (AUC values between 0.41–0.74) was lower compared to the radiomics models (AUC values between 0.48–0.89) in the training cohorts. The validation results of both models showed a wider range of diagnostic performance parameters compared to the training results possibly explained by the small dataset, the methodology used for selection and model building, and potential overfitting. The wide AUC range in the clinical models leads us to the hypothesis that the small dataset contains unseen biological covariates, and that therefore the wide AUC range in the radiomics models cannot be explained by variations in imaging alone.

To the best of our knowledge, this is the first study investigating the role of MRI-based radiomics for the prediction of axillary lymph node metastasis in breast cancer patients by extracting features from delineated axillary lymph nodes. Previously published articles investigated the same topic by extracting the features from the delineated breast tumor [15,20,21]. These articles showed promising validation results with AUC values between 0.77–0.82. In this recent study, initially, the small ROI volumes were seen as a reason for the inconclusive results. If an ROI contains a low number of voxels, it may not be possible to calculate meaningful radiomics features [22]. However, after the subanalysis excluding ROI volumes less than 50 voxels, the AUC values were between 0.53–0.82 and 0.41–0.83 for the training cohorts for models 1a and 2a, respectively, which highlights the effects of differences in scan acquisition and reconstruction parameters. Furthermore, the skewed data in this recent study may have caused inconsistent results compared to the previous studies as models tend to favor the more common outcome.

To date, only two previously published articles extracted features from delineated lymph nodes for radiomics and deep learning analyses. The first article used a neural network

to develop prediction models in head and neck cancer ^[23]. The second article developed a radiomics model based on CT images of colorectal cancer patients ^[24]. Both studies showed that there is potential by delineating lymph nodes for radiomics and deep learning analysis for the classification of positive and negative lymph nodes. The differences in results compared to this recent study may be due to the variety of implementation of the different steps in the radiomics workflow and the chosen imaging modality (CT vs. MRI).

The diagnostic performance of dedicated axillary T2W MRI for axillary lymph node staging has previously been investigated using node-by-node analysis ^[12]. Schipper et al. showed AUC values between 0.78–0.88, with a good interobserver agreement ($\kappa = 0.70$). The current analysis with MRI-based radiomics using dedicated axillary T2W MR images suggested that the quantitative analysis did not exceed the qualitative analysis by the radiologists. It was decided to only perform radiomics analyses using the T2W MR images, as previous research indicated that diffusion-weighted images and apparent diffusion coefficient measurements have no added value for the axillary lymph node staging ^[12,25]. Furthermore, a recently published article has shown that the evaluation of axillary lymph nodes with dedicated axillary MRI is comparable to standard breast MRI with a complete field of view of the axillary region ^[25]. However, the majority of the breast MRI examinations are still performed with an incomplete field of view of the axillary region ^[9]. In addition, the coronal view of the dedicated axillary MRI possibly provides more accurate delineations compared to the transversal view of the standard breast MRI, which could be of added value to the radiomics analysis.

Most radiomics studies suffer from small and heterogeneous datasets collected from different imaging systems. In this current study, a great advantage for the radiomics analyses was the prospectively collected set of MR images on the same MRI scanner using an equal acquisition protocol with the patients in corresponding positions. Despite the prospectively collected dataset, a number of acquisition and reconstruction parameters varied depending on the patient. Furthermore, the different sets of features selected in every training cohort resulted in a wide range of AUC values and did not allow the development of a final radiomics prediction model. This could be justified by two theories: (i) The variations in acquisition and reconstruction parameters significantly affected the value of radiomics features, resulting in non-comparable data points; or (ii) Radiomics features do not have an added value in the prediction of axillary lymph nodes metastasis. However, theory (ii) is less likely, as radiomics models performed well in some splits. Future MRI phantom and reproducibility studies should investigate the effect of MR image acquisition and reconstruction parameters on feature values to determine repeatable and reproducible features. We nevertheless believe that it is also important to publish inconclusive radiomics results since publication bias seems to play a role in this research field, with only 6% of the radiomics articles presenting negative results ^[26].

This study also has certain limitations. The large skewness of the data with only 7% positive axillary lymph nodes was a drawback for the analyses. The skewness of the data was addressed by splitting the dataset using two different strategies and by using repeated cross-validation in the training cohort. However, it is important to note that the ratio of node-positive (19%) and node-negative (81%) breast cancer patients in this study is comparable to the clinics. Besides the skewness of the data, the included number of patients was relatively low for radiomics analysis and selecting only node-positive patients in strategy 2 decreased the number even further. However, since the dedicated axillary MRI is not included in the breast MRI protocol and no similar public dataset is available, it is not possible to expand this current dataset. Lastly, manual delineation of the axillary lymph nodes was performed by one researcher, which potentially could be a major limitation of the findings because of the susceptibility of inter- and intra-observer variabilities ^[27]. Although this issue has been addressed in this current study by developing models based on only robust features for varying breast tumor delineations ^[28]. Based on the assumption that breast and lymph node delineations on MRI are comparable, varying delineations did not affect the radiomics results. However, this topic needs to be thoroughly investigated in future studies.

Materials and Methods

Patient Population

Consecutive women with histopathologically proven breast cancer, who had undergone dedicated axillary MRI between August 2012 and October 2014, followed by sentinel lymph node biopsy (SNLB) or axillary lymph node dissection (ALND), were considered for inclusion. Patients were excluded if they had undergone neoadjuvant systemic therapy before axillary surgery and in the case of ductal carcinoma in situ only. This study was approved by the local medical ethics committee, and the requirement of written informed consent was waived due to the retrospective study design. Fifty of the dedicated axillary T2W and diffusion-weighted MR images were earlier described by Schipper et al. for axillary lymph node staging, and 90 of the dedicated axillary T2W and gadofosveset-enhanced MR images were earlier described by Van Nijnatten et al. for axillary lymph node staging ^[12,29].

Clinical and Pathological Characteristics

Clinical and pathological data were derived from the patients' medical records: age, clinical TNM stage, pathological TNM stage, tumor histology, tumor grade, breast cancer subtype, and type of axillary surgery. Lymph nodes with isolated tumor cells (≤ 0.2 mm) and micrometastases (>0.2 – ≤ 2.0 mm) were considered negative, while those with macrometastases (>2.0 mm) were considered positive.

MRI Acquisition

The dedicated axillary MR images were performed using a 32-channel cardiac coil on a 3.0 Tesla scanner (Achieva, Philips Healthcare, Best, the Netherlands). During the MRI examination, the patient was positioned in a supine position with the ipsilateral arm elevated. The anatomical confines of the dedicated axillary MR images were between the humeral head and the inferior border of the scapula. The MRI protocol included an unenhanced three-dimensional T2W turbo spin-echo sequence without fat suppression (pixel size, 1.25 × 1.25 mm; repetition time, 2000 ms; echo time between 150–202 ms; echo train length, 52 or 66; flip angle, 90°; acquisition slice thickness, 2.5 mm; reconstruction slice thickness, 1.25 mm), a contrast-enhanced T1-weighted sequence, and a diffusion-weighted imaging sequence with fat suppression.

MRI Lymph Node Delineation

All axillary lymph nodes of each dedicated axillary T2W MR image were manually delineated in three dimensions using MIM software (version 6.9.4, MIM Software Inc., Cleveland, OH, USA) by a medical researcher (S.S.) with three years of experience in axillary lymph node imaging validated by a dedicated breast radiologist (M.L.) with eleven years of experience (Figure 4). No clinical information and pathology results were available during delineation and validation. The delineated lymph nodes were subsequently matched with their histopathological findings (node-by-node matching). Reliable node-by-node matching was obtained using single-photon emission computed tomography-X-ray computed tomography (SPECT-CT) in patients undergoing SLNB, and an anatomical map was used for patients undergoing ALND. The exact procedure of the node-by-node matching was previously described by Schipper et al. [30].

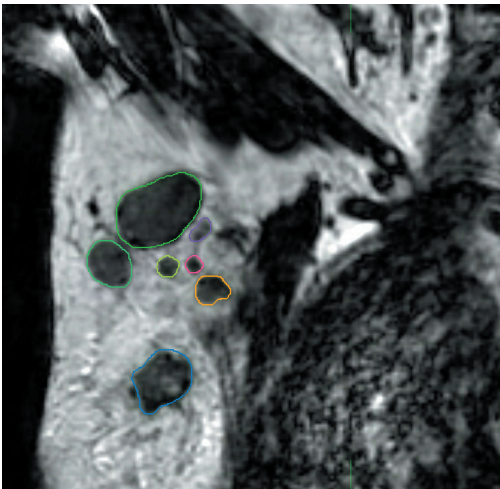


Figure 4: Coronal T2-weighted dedicated axillary MR image of a 55-year old woman with invasive breast cancer, who was treated with mastectomy and axillary lymph node dissection (pT1N2). The MR image demonstrates an example of delineations of lymph nodes in the right axilla on the MIM software.

MRI Preprocessing and Feature Extraction

Image preprocessing of the T2W images was performed after delineation. Bias field correction was applied to every T2W MR image using MIM software to correct for non-uniform grayscale intensities caused by field inhomogeneities. To ensure better comparability of voxel intensities, additional image normalization and discretization was performed by the open-source Pyradiomics software (version 2.2.0) prior to feature extraction [31]. For discretization, grayscale values were aggregated with a fixed bin width of 10, which ensured the recommended amount of bins between 30–130 [31]. Resampling was not required, as all images consisted of isotropic voxels of equal size 1.25 mm³. Quantitative radiomics features were extracted from the delineated regions using the Pyradiomics software. The extracted features can be subdivided into the following classes: first-order statistics, three-dimensional shape-based, gray level co-occurrence matrix, gray level run length matrix, gray level size zone matrix, neighboring gray-tone difference matrix, and gray level dependence matrix.

Radiomics Feature Selection and Model Development

Taking into account the small skewed dataset and the unavailability of an external validation dataset, the data were randomly divided into training and validation cohort 100 times using two different strategies to create a more balanced training cohort. In the first strategy, 85% (12 out of 14) of the node-positive (i.e., patients with axillary lymph node metastasis at final pathology) breast cancer patients were selected in the training cohort, and all remaining node-positive and node-negative (i.e., patients without axillary lymph node metastasis at final pathology) patients in the validation cohort, considering each axillary lymph node as an individual data point when training the model. In the second strategy, only the lymph nodes of patients with node-positive breast cancer were considered as individual data points when training and validating the model. To maintain the original class imbalance of the node-positive patients, 10 patients were selected in the training cohort. For both strategies, additional models were developed using a random undersampled balanced training cohort. All lymph nodes of one patient were always included in either the training cohort or the validation cohort, and therefore each split caused a varying number of positive lymph nodes in each cohort. Feature selection started with the removal of near-zero variance features followed by the removal of highly correlated features using the Pearson pairwise correlation greater than 0.95. Subsequently, recursive feature elimination with bagged trees was applied with repeated 5-fold cross-validation to select a maximum number of features in the training cohort. The number of features was chosen at the point when the addition of more features did not increase the diagnostic performance of the models. Random forest binary classification models were trained, using optimized random forest parameters (number of trees and features per node) for the training cohort, selecting the optimal number of features for each generated model. In addition, a separate set of models was

generated using the same pipeline but by adding an additional feature selection step at the very beginning. In this step, features robust to the variability of manual delineations of breast tumors on MRI by four observers were selected according to three different cut-off values (intraclass correlation coefficient of >0.75 , >0.80 , and >0.90) [28]. Figure 5 provides an overview of strategies 1 and 2 with the different developed models.

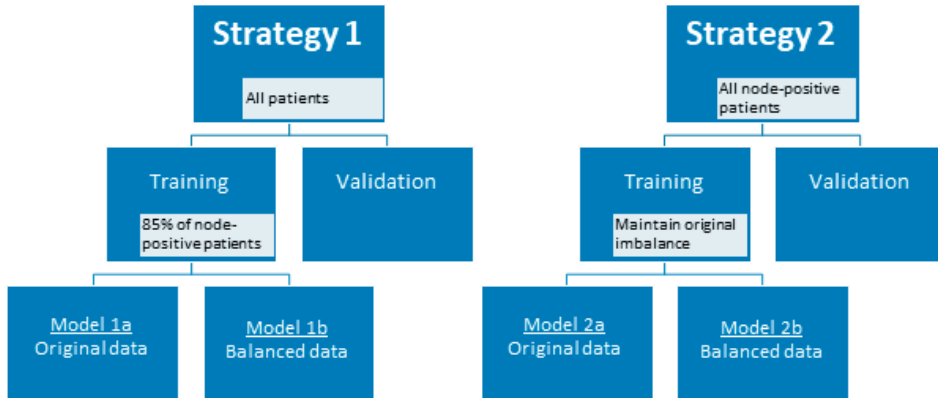


Figure 5: Model strategies.

Radiomics Subanalysis

A separate set of models was generated using the first and second strategies as described earlier on a dataset where ROIs with less than 50 voxels were excluded [31]. On these models, only the additional feature selection step with different intraclass correlation coefficient cut-off values was not performed.

Clinical Model Development

Clinical models were trained based on clinical characteristics available before the axillary surgery. Random forest models with bagged tree function for the prediction of axillary lymph node metastasis were trained and validated using the same strategies as described above, except for the feature selection step, which was only the removal of highly correlated clinical characteristics. These clinical models were used to indicate the effect of known and unknown patient's biological covariates compared to a pure imaging-based model as well as to rank the importance of the clinical characteristics in this dataset using the Gini impurity method.

Statistical Analyses and Study Evaluation

The statistical analyses, including dataset splitting and balancing, feature selection, model development, and performance evaluation, were performed in R (version 3.6.3; <http://www.r-project.org>) using R studio (version 1.2.1335, Vienna, Austria) [32]. The performance of all models was assessed using the area under the receiver operating

characteristics curve (AUC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). The Spearman correlation was used to calculate the correlation between the number of voxels per ROI and the corresponding pathological outcome. The radiomics workflow was evaluated using the radiomics quality score (RQS) [33]. This study followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines [34].

Conclusions

In conclusion, based on our results dedicated axillary MRI-based radiomics with node-by-node analysis did not contribute to the prediction of axillary lymph node metastasis based on data where variations in acquisition and reconstruction parameters were not addressed. Larger datasets combined with MRI phantom data and reproducibility studies are necessary to determine if further radiomics research using dedicated axillary MR images for the prediction of axillary lymph node metastasis is of added value.

Supplementary Materials

The following are available online at www.mdpi.com/xxx/s1, Supplementary Material A includes a list of how often each feature was chosen in the 100 iterations for each model (excel file). Supplementary Material B includes Figure S1: Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature selection step ($ICC > 0.75$): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort, Figure S2: Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature selection step ($ICC > 0.80$): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort, Figure S3: Violin plots for the radiomics models developed using the first (A) and second (B) strategy with additional feature selection step ($ICC > 0.90$): AUC value distributions (100 iterations) for the four models (1a, 1b, 2a, and 2b) in both the training and validation cohort, Figure S4: Violin plots for the radiomics models with the exclusion of ROIs < 50 voxels developed using the first strategy and second strategy: AUC value distribution (100 iterations) for the two models (1a and 2a) in both the training and validation cohort, Table S1: the diagnostic performance of the radiomics models (100 iterations) with the exclusion of ROIs < 50 voxels for the first and second strategy, Table S2: Radiomics Quality Score, Table S3: TRIPOD Checklist.

Author Contributions

Conceptualization, S.S. and R.W.Y.G.; methodology, S.S. and R.W.Y.G.; software, S.P.; validation, S.S., R.W.Y.G., and A.I.; formal analysis, S.S., R.W.Y.G., A.I., and H.C.W.; investigation, S.S. and R.W.Y.G.; resources, S.S. and T.J.A.v.N.; data curation, S.S., T.J.A.v.N., and M.B.I.L.; writing—original draft preparation, S.S. and R.W.Y.G.; writing—review and editing, S.S., R.W.Y.G., A.I., S.P., M.B.I.L., R.G.H.B.-T., T.J.A.v.N., S.M.E.E., H.C.W., and M.L.S.; visualization, S.S., and R.W.Y.G.; supervision, M.L.S. and H.C.W.; project administration, S.S. and R.W.Y.G.; funding acquisition, M.L.S. All authors have read and agreed to the published version of the manuscript.

Funding

S. Samiei received a salary from the Alpe d’HuZes Foundation (Dutch Cancer Society; grant number UM 2013-6229).

Institutional Review Board Statement

The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of the University Hospital Maastricht and Maastricht University (METC azM / UM) (Ethic code: 2017-0199 with date of approval 12 October 2017).

Informed Consent Statement

Patient consent was waived due to the retrospective design of the study.

Data Availability Statement

The data presented in this study are available on reasonable request from the corresponding author. Due to privacy restrictions the data are not publicly available.

Conflicts of Interest

The authors declare no conflict of interest

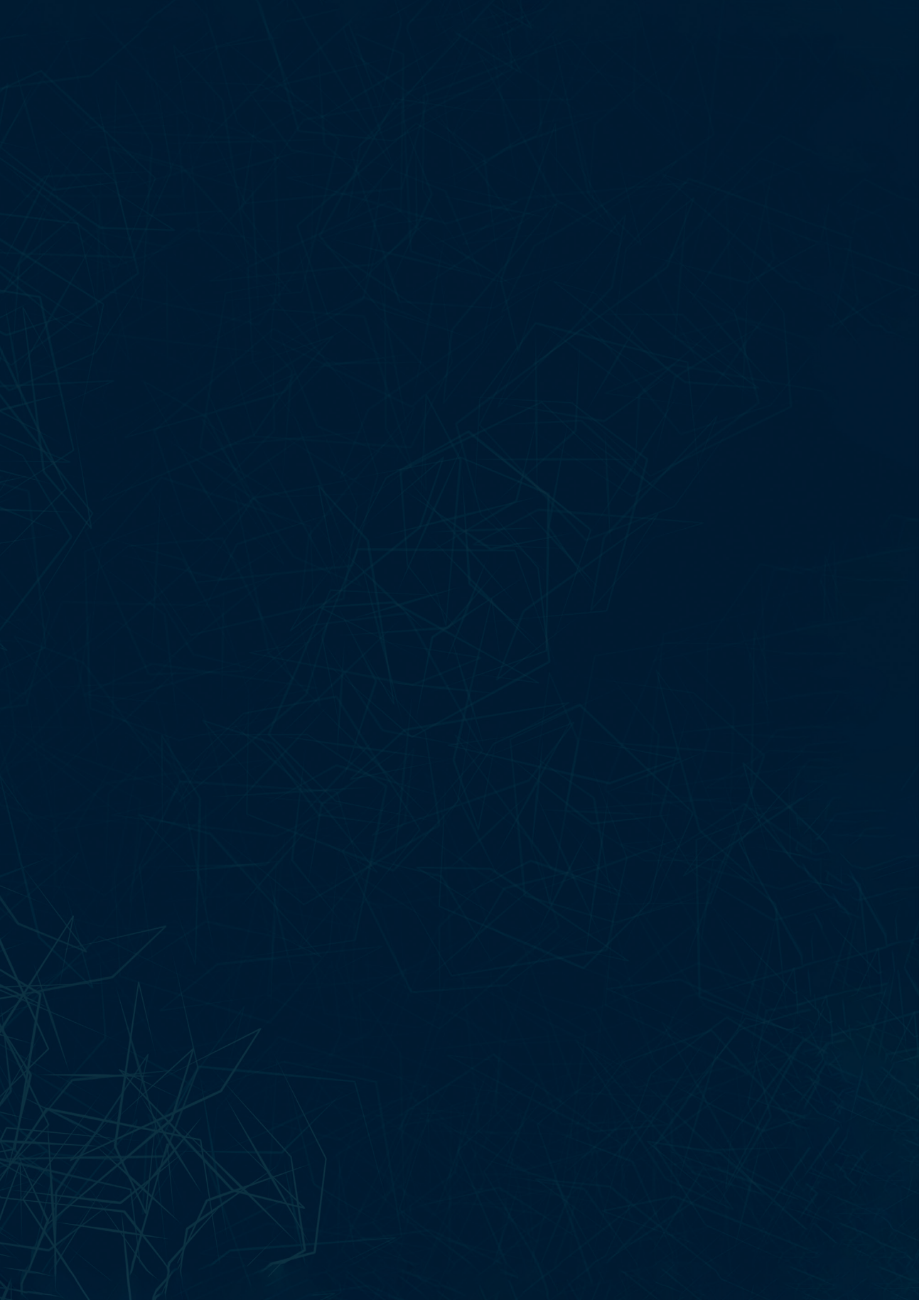
References

1. Beenken, S.W.; Urist, M.M.; Zhang, Y.; Desmond, R.; Krontiras, H.; Medina, H.; Bland, K.I. Axillary lymph node status, but not tumor size, predicts locoregional recurrence and overall survival after mastectomy for breast cancer. *Ann. Surg.* **2003**, *237*, 732–738; discussion 738–739, doi:10.1097/01.SLA.0000065289.06765.71.
2. Soerjomataram, I.; Louwman, M.W.; Ribot, J.G.; Roukema, J.A.; Coebergh, J.W. An overview of prognostic factors for long-term survivors of breast cancer. *Breast Cancer Res. Treat.* **2008**, *107*, 309–330, doi:10.1007/s10549-007-9556-1.
3. Carter, C.L.; Allen, C.; Henson, D.E. Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer* **1989**, *63*, 181–187, doi:10.1002/1097-0142(19890101)63:1<181::aid-cncr2820630129>3.0.co;2-h.
4. Fisher, B.; Bauer, M.; Wickerham, D.L.; Redmond, C.K.; Fisher, E.R.; Cruz, A.B.; Foster, R.; Gardner, B.; Lerner, H.; Margolese, R.; et al. Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update. *Cancer* **1983**, *52*, 1551–1557, doi:10.1002/1097-0142(19831101)52:9<1551::aid-cncr2820520902>3.0.co;2-3.
5. Surveillance, Epidemiology, and End Results Program (SEER). Table 4.13: Cancer of the Female Breast (Invasive). 5-Year Relative and Period Survival by Race, Diagnosis Year, Age and Stage at Diagnosis. In: SEER Cancer Statistics Review (CSR) 1975-2012. Available online: https://seer.cancer.gov/archive/csr/1975_2012/browse_csr.php?sectionSEL=4&pageSEL=sect_04_table.13.html (accessed on 31 May 2020).
6. Senkus, E.; Kyriakides, S.; Ohno, S.; Penault-Llorca, F.; Poortmans, P.; Rutgers, E.; Zackrisson, S.; Cardoso, F.; Committee, E.G. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **2015**, *26* (Suppl. S5), v8–v30, doi:10.1093/annonc/mdv298.
7. Caudle, A.S.; Cupp, J.A.; Kuerer, H.M. Management of axillary disease. *Surg. Oncol. Clin. N. Am.* **2014**, *23*, 473–486, doi:10.1016/j.soc.2014.03.007.
8. Sardanelli, F.; Boetes, C.; Borisch, B.; Decker, T.; Federico, M.; Gilbert, F.J.; Helbich, T.; Heywang-Kobrunner, S.H.; Kaiser, W.A.; Kerin, M.J.; et al. Magnetic resonance imaging of the breast: Recommendations from the EUSOMA working group. *Eur. J. Cancer* **2010**, *46*, 1296–1316, doi:10.1016/j.ejca.2010.02.015.
9. van Nijnatten, T.J.A.; Ploumen, E.H.; Schipper, R.J.; Goorts, B.; Andriessen, E.H.; Vanwetswinkel, S.; Schavemaker, M.; Nelemans, P.; de Vries, B.; Beets-Tan, R.G.H.; et al. Routine use of standard breast MRI compared to axillary ultrasound for differentiating between no, limited and advanced axillary nodal disease in newly diagnosed breast cancer patients. *Eur. J. Radiol.* **2016**, *85*, 2288–2294, doi:10.1016/j.ejrad.2016.10.030.
10. Kvistad, K.A.; Rydland, J.; Smethurst, H.B.; Lundgren, S.; Fjosne, H.E.; Haraldseth, O. Axillary lymph node metastases in breast cancer: Preoperative detection with dynamic contrast-enhanced MRI. *Eur. Radiol.* **2000**, *10*, 1464–1471, doi:10.1007/s003300000370.
11. Murray, A.D.; Staff, R.T.; Redpath, T.W.; Gilbert, F.J.; Ah-See, A.K.; Brookes, J.A.; Miller, I.D.; Payne, S. Dynamic contrast enhanced MRI of the axilla in women with breast cancer: Comparison

- with pathology of excised nodes. *Br. J. Radiol.* **2002**, *75*, 220–228, doi:10.1259/bjr.75.891.750220.
12. Schipper, R.J.; Paiman, M.L.; Beets-Tan, R.G.; Nelemans, P.J.; de Vries, B.; Heuts, E.M.; van de Vijver, K.K.; Keymeulen, K.B.; Brans, B.; Smidt, M.L.; et al. Diagnostic Performance of Dedicated Axillary T2- and Diffusion-weighted MR Imaging for Nodal Staging in Breast Cancer. *Radiology* **2015**, *275*, 345–355, doi:10.1148/radiol.14141167.
 13. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* **2016**, *278*, 563–577, doi:10.1148/radiol.2015151169.
 14. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.; Granton, P.; Zegers, C.M.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **2012**, *48*, 441–446, doi:10.1016/j.ejca.2011.11.036.
 15. Dong, Y.; Feng, Q.; Yang, W.; Lu, Z.; Deng, C.; Zhang, L.; Lian, Z.; Liu, J.; Luo, X.; Pei, S.; et al. Preoperative prediction of sentinel lymph node metastasis in breast cancer based on radiomics of T2-weighted fat-suppression and diffusion-weighted MRI. *Eur. Radiol.* **2018**, *28*, 582–591, doi:10.1007/s00330-017-5005-7.
 16. Han, L.; Zhu, Y.; Liu, Z.; Yu, T.; He, C.; Jiang, W.; Kan, Y.; Dong, D.; Tian, J.; Luo, Y. Radiomic nomogram for prediction of axillary lymph node metastasis in breast cancer. *Eur. Radiol.* **2019**, *29*, 3820–3829, doi:10.1007/s00330-018-5981-2.
 17. Yang, J.B.; Wang, T.; Yang, L.F.; Wang, Y.B.; Li, H.M.; Zhou, X.B.; Zhao, W.L.; Ren, J.C.; Li, X.Y.; Tian, J.; et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. *Sci. Rep.* **2019**, *9*, 1–11, doi:10.1038/s41598-019-40831-z.
 18. Yu, F.H.; Wang, J.X.; Ye, X.H.; Deng, J.; Hang, J.; Yang, B. Ultrasound-based radiomics nomogram: A potential biomarker to predict axillary lymph node metastasis in early-stage invasive breast cancer. *Eur. J. Radiol.* **2019**, *119*, 108658, doi:10.1016/j.ejrad.2019.108658.
 19. Liu, C.; Ding, J.; Spuhler, K.; Gao, Y.; Serrano Sosa, M.; Moriarty, M.; Hussain, S.; He, X.; Liang, C.; Huang, C. Preoperative prediction of sentinel lymph node metastasis in breast cancer by radiomic signatures from dynamic contrast-enhanced MRI. *J. Magn. Reson. Imaging* **2019**, *49*, 131–140, doi:10.1002/jmri.26224.
 20. Chai, R.M.; Ma, H.; Xu, M.J.; Arefan, D.; Cui, X.Y.; Liu, Y.; Zhang, L.N.; Wu, S.D.; Xu, K. Differentiating axillary lymph node metastasis in invasive breast cancer patients: A comparison of radiomic signatures from multiparametric breast MR sequences. *J. Magn. Reson. Imaging* **2019**, *50*, 1125–1132, doi:10.1002/jmri.26701.
 21. Tan, H.N.; Gan, F.W.; Wu, Y.P.; Zhou, J.; Tian, J.; Lin, Y.S.; Wang, M.Y. Preoperative Prediction of Ancillary Lymph Node Metastasis in Breast Carcinoma Using Radiomics Features Based on the Fat-Suppressed T2 Sequence. *Acad. Radiol.* **2020**, *27*, 1217–1225, doi:10.1016/j.acra.2019.11.004.
 22. Court, L.E.; Fave, X.; Mackin, D.; Lee, J.; Yang, J.Z.; Zhang, L.F. Computational resources for radiomics. *Transl. Cancer Res.* **2016**, *5*, 340–348, doi:10.21037/tcr.2016.06.17.
 23. Ho, T.Y.; Chao, C.H.; Chin, S.C.; Ng, S.H.; Kang, C.J.; Tsang, N.M. Classifying Neck Lymph Nodes of Head and Neck Squamous Cell Carcinoma in MRI Images with Radiomic Features. *J. Digit.*

Chapter 4

- Imaging* **2020**, *33*, 613–618, doi:10.1007/s10278-019-00309-w.
24. Li, M.; Zhang, J.; Dan, Y.; Yao, Y.; Dai, W.; Cai, G.; Yang, G.; Tong, T. A clinical-radiomics nomogram for the preoperative prediction of lymph node metastasis in colorectal cancer. *J. Transl. Med.* **2020**, *18*, 46, doi:10.1186/s12967-020-02215-0.
 25. Samiei, S.; Smidt, M.L.; Vanwetswinkel, S.; Engelen, S.M.E.; Schipper, R.J.; Lobbes, M.B.I.; van Nijnatten, T.J.A. Diagnostic performance of standard breast MRI compared to dedicated axillary MRI for assessment of node-negative and node-positive breast cancer. *Eur. Radiol.* **2020**, *30*, 4212–4222, doi:10.1007/s00330-020-06760-6.
 26. Buvat, I.; Orlhac, F. The Dark Side of Radiomics: On the Paramount Importance of Publishing Negative Results. *J. Nucl. Med.* **2019**, *60*, 1543–1544, doi:10.2967/jnumed.119.235325.
 27. Saha, A.; Grimm, L.J.; Harowicz, M.; Ghate, S.V.; Kim, C.; Walsh, R.; Mazurowski, M.A. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med. Phys.* **2016**, *43*, 4558, doi:10.1118/1.4955435.
 28. Granzier, R.W.Y.; Verbakel, N.M.H.; Ibrahim, A.; van Timmeren, J.E.; van Nijnatten, T.J.A.; Leijenaar, R.T.H.; Lobbes, M.B.I.; Smidt, M.L.; Woodruff, H.C. MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Sci. Rep.* **2020**, *10*, 14163, doi:10.1038/s41598-020-70940-z.
 29. van Nijnatten, T.J.A.; Schipper, R.J.; Lobbes, M.B.I.; van Roozendaal, L.M.; Voo, S.; Moosdorff, M.; Paiman, M.L.; de Vries, B.; Keymeulen, K.; Wildberger, J.E.; et al. Diagnostic performance of gadofosveset-enhanced axillary MRI for nodal (re)staging in breast cancer patients: Results of a validation study. *Clin. Radiol.* **2018**, *73*, 168–175, doi:10.1016/j.crad.2017.09.005.
 30. Schipper, R.J.; Smidt, M.L.; van Roozendaal, L.M.; Castro, C.J.; de Vries, B.; Heuts, E.M.; Keymeulen, K.B.; Wildberger, J.E.; Lobbes, M.B.; Beets-Tan, R.G. Noninvasive nodal staging in patients with breast cancer using gadofosveset-enhanced magnetic resonance imaging: A feasibility study. *Investig. Radiol.* **2013**, *48*, 134–139, doi:10.1097/RLI.0b013e318277f056.
 31. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.C.; Pieper, S.; Aerts, H. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* **2017**, *77*, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
 32. Racine, J.S. RStudio: A platform-independent IDE for R and Sweave. *J. Appl. Econom.* **2012**, *27*, 167–172.
 33. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.; Even, A.J.G.; Jochems, A.; et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **2017**, *14*, 749–762, doi:10.1038/nrclinonc.2017.141.
 34. Moons, K.G.; Altman, D.G.; Reitsma, J.B.; Ioannidis, J.P.; Macaskill, P.; Steyerberg, E.W.; Vickers, A.J.; Ransohoff, D.F.; Collins, G.S. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann. Intern. Med.* **2015**, *162*, W1–W73, doi:10.7326/M14-0698



PART III

A large, white, stylized number 5 is centered on a blue, textured, watercolor-like background. The background consists of various shades of blue, from light to dark, with a mottled, organic appearance. The number 5 is a simple, bold, sans-serif font. The overall composition is abstract and artistic.

5

Chapter 5

Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods

Authors

Shruti Atul Mali, Abdalla Ibrahim, Henry C. Woodruff, Vincent Andrearczyk,
Henning Müller, Sergey Primakov, Zohaib Salahuddin, Avishek Chatterjee
and Philippe Lambin

Adapted from

Journal of Personalized Medicine. 2021 Sep;11(9):842

DOI

10.3390/jpm11090842

Abstract

Radiomics converts medical images into mineable data via a high-throughput extraction of quantitative features used for clinical decision support. However, these radiomic features are susceptible to variation across scanners, acquisition protocols, and reconstruction settings. Various investigations have assessed the reproducibility and validation of radiomic features across these discrepancies. In this narrative review, we combine systematic keyword searches with prior domain knowledge to discuss various harmonization solutions to make the radiomic features more reproducible across various scanners and protocol settings. Different harmonization solutions are discussed and divided into two main categories: image domain and feature domain. The image domain category comprises methods such as the standardization of image acquisition, post-processing of raw sensor-level image data, data augmentation techniques, and style transfer. The feature domain category consists of methods such as the identification of reproducible features and normalization techniques such as statistical normalization, intensity harmonization, ComBat and its derivatives, and normalization using deep learning. We also reflect upon the importance of deep learning solutions for addressing variability across multi-centric radiomic studies especially using generative adversarial networks (GANs), neural style transfer (NST) techniques, or a combination of both. We cover a broader range of methods especially GANs and NST methods in more detail than previous reviews.

Keywords

radiomics; harmonization; feature reproducibility; deep learning; medical imaging

Introduction

Medical imaging is routinely used in clinical practice to assist the decision-making process for diagnostic and treatment purposes [1,2]. Radiomics is an emerging field within medical image analysis that goes beyond qualitative assessment by extracting a large number of quantitative image features [3,4]. The radiomic hypothesis postulates that the quantitative study of medical image data can provide complementary knowledge in a quick and reproducible manner to support clinicians in their decision-making process, assisted by automated or semi-automated software [5,6]. The information acquired can help advance the clinical decision support systems to connect the link between radiomic features and clinical endpoints by building diagnostic, prognostic, and predictive analysis models. Radiomics is the consequence of many decades of computerized diagnosis, prognosis, and treatment research [7,8]. A powerful radiomics approach involves the extraction of various quantitative features from medical images, storing this data in a federated form of a database [9] where several individual databases function as an entity, and the successive mining of data to acquire relevant clinical outcomes [10]. Large quantities of data are required to develop robust predictive models and this amount of data is usually obtained from multiple hospitals and/or institutions. Furthermore, due to the continuous improvement in scanner and protocol settings, this type of data is a moving target. To compensate for the effects scanner/protocol variability might have on the predictive models, large quantities of data are needed to make systems generalize. In these cases, federated (or distributed) learning could be adapted to allow sharing of data between hospitals/institutes to develop robust predictive models [10]. Major management problems still exist even though there are databases that are collecting and cross-referencing massive amounts of radiomics information in addition to other related patient data from millions of case studies [11–14].

Radiomic feature extraction can be categorized into two main approaches: hand-crafted (derived from traditional statistical and computer vision methods) and deep learning (DL). Hand-crafted radiomics characteristics (such as texture, shape, intensity) provide information on the particular area of the medical imaging scan, often referred to as the region or volume of interest (ROI or VOI), which could be a tumor, a tissue, or an organ as a whole [15]. DL is also a data-driven method that is inspired by the biological neural networks in the human brain. The difference between hand-crafted and DL approaches mostly lies in the way visual representations are learned. For example, some DL algorithms learn complex visual features and perform ROI segmentation using cascading layers with non-linearities by using ‘sliding’ kernels in convolutional neural networks (CNN), while hand-crafted features represent the spatial appearances (texture and shape) by mathematically extracting spatial distribution on inter-pixel relationships, signal intensities, gray-scale patterns, and spectral properties [16]. DL has the benefit

of not necessarily requiring prior segmentation masks of the medical imaging scan. However, DL is a 'black box' approach, i.e., the lack of interpretability of the models and the deep features generated are seen as a key limitation in clinical applications [17]. DL also re-quires a larger amount of data and/or pre-trained models often trained on diverse do-mains (e.g., photographic images), in order to perform efficiently and effectively. The vast majority of published radiomic models lack consistent evaluation of performance, suffi-cient large-scale annotated datasets for radiomic studies, reproducibility, clinical efficacy, and large-scale validation on sufficiently large cohorts, despite these being prerequisites for clinical translation [18,19]. Furthermore, there is a lack of reproducibility of radiomic features while translating results into clinical practice [20]. Ideally, the features extracted using radiomics represent imaging biomarkers and should be independent of image acquisition parameters or protocols [21]. For example, if a patient is scanned in different hospitals, the quantitative features extracted from all these scans should either have similar values or the correct transformation should be known. Scanner protocols and hardware are constantly changing over time and differ across hospitals. The same scanner can also be configured differently. Frequent software updates might have an influence on images produced. A major consequence of these scanner and protocol variations is a domain shift [22], i.e., a shift in data distribution across various cen-ters/time/machines/software. Please see Figure 1 showing inter-center variation in data distribution obtained from PET/CT scans from HEad and neCK TumOR (HECKTOR) challenge [23].

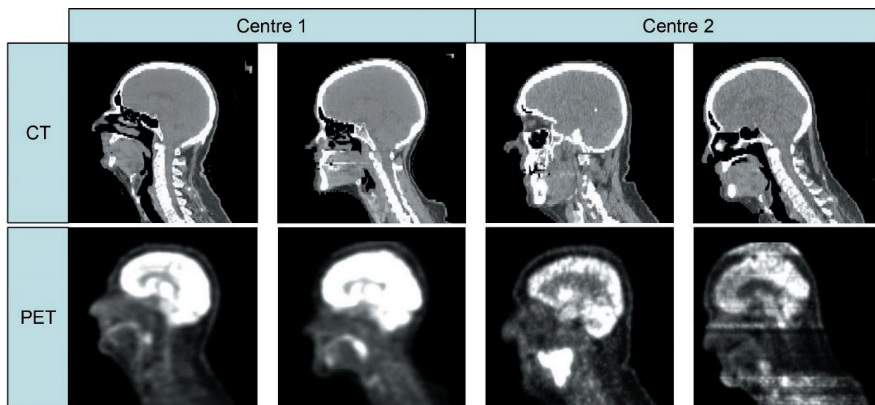


Figure 1: PET and CT slices obtained from two different centers (Center 1 = Centre Hospitalier Universitaire de Sherbrooke, Sherbrooke, Canada and Center 2 = Hôpital Maisonneuve-Rosemont, Montréal, Canada). The top row shows CT images while the bottom row shows PET images. The four columns indicate four different patients. Adapted from [23].

Studies have shown the effects of image acquisition parameters on the reproducibility of radiomic features [21,24–27]. Many studies [1,28–31] have also explored the discriminative power of radiomic features. However, the reproducibility of a radiomic feature

does not guarantee its discriminative power [32,33], and thus the two aspects of reproducibility and discriminative power cannot be treated in isolation. For instance, a feature may have excellent reproducibility across scanner and protocol variations but have no discriminative power for the problem of interest. The scanner and/or protocol variability could hamper the stability as well as the discriminative power of the features. Feature variability is also caused due to varying contours or ROIs. For example, Yang et al. [34] observed that gray-level neighborhood difference matrices (GLNDM) based radiomic features were most robust against the manual contouring variability in PET scans of lung cancer. Variation in inter-observer delineation has an impact on radiomic analysis and is examined in [34–38]. These variations can have repercussions on image texture and consequently on the radiomic features. Different feature extraction algorithms and image processing techniques also influence the feature variation and have been addressed by the image biomarker standardization initiative [39]. However, in this work, we only focus on studies that investigate radiomic feature reproducibility across scanner and protocol variations. Various methods have been proposed in the literature to improve the re-reproducibility of radiomic features across scanner and protocol variations and a few of these harmonization methods have been reviewed in [40,41]. In addition to feature robustness, investigations should be carried out to ascertain model accuracy/performance as well. For instance, the model should have sufficient data to achieve predictive performance at least equal to the current clinical standard; the model should be externally and/or internally validated across different centers; several performance metrics such as the area under curve (AUC) of the receiver operating characteristic (ROC) curve and precision recall (PR) curves can be used to evaluate the model performance.

The organization of the paper is as follows. We primarily categorize the methods under the image domain or the feature domain. The harmonization methods discussed under the image domain (Section 3) are performed on the whole image (raw or reconstructed) before feature extraction and thus aim to harmonize images acquired across different centers/scanners/protocols. In this section, we briefly review methods in such a way that they can be applied at every stage of medical image processing from image acquisition to image analysis (Figure 2). This section starts with a discussion on various standards for image acquisition/reconstruction parameters. Moving forward, post-processing methods for raw sensor-level image data followed by brief reviews of existing image analysis techniques (e.g., data augmentation techniques using generative adversarial networks (GANs) and style transfer) are discussed. The methods categorized under the feature domain (Section 4) are performed after (or within) feature extraction and aim to harmonize extracted radiomic features. In this section, the methods are listed in order of their complexity. Under the feature domain, we briefly review two approaches: identification of reproducible features (a convenient approach) and normalization techniques (statistical approaches). The normalization techniques are further divided into

basic statistical normalization (rescaling/standardization); intensity harmonization techniques; ComBat method and its derivatives; normalization using DL. The overall objective of this review is to address the advantages, disadvantages, and challenges posed by these harmonization methods. Figure 2 shows an overview of different harmonization methods that are applicable at different stages of medical imaging.

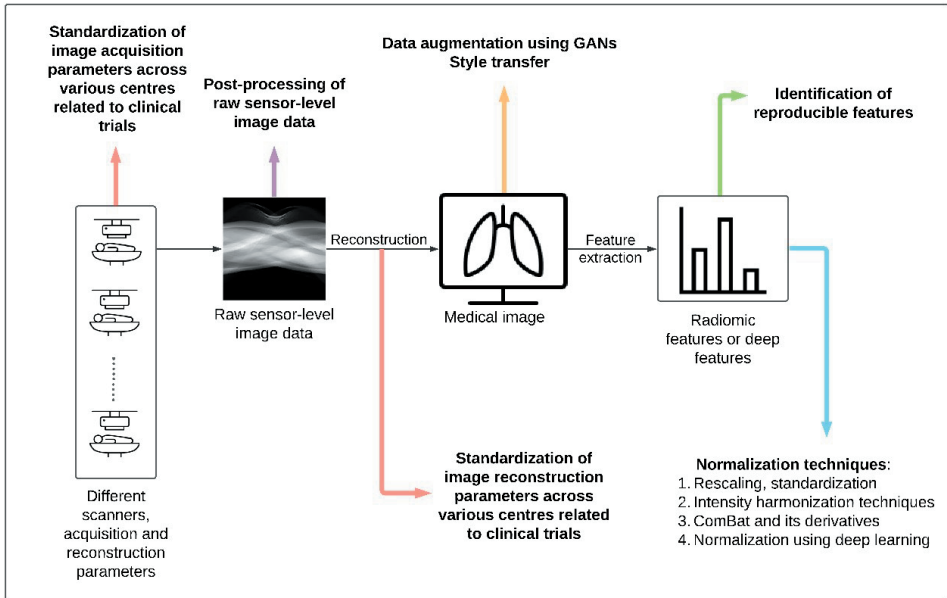


Figure 2: Overview of harmonization methods at different stages of medical imaging.

Search strategy: Our search strategy for this review was based on a set of research questions:

1. Have scanner and protocol variations affected the reproducibility of radiomic features/images? If yes, the how significant was the change?
2. Various harmonization methods were identified in previous work. Can they be categorized into domains (image and feature)? Furthermore, can the methods be applied at different stages of medical imaging (Figure 2)?
3. What are the latest developments in the field of radiomics to make radiomics more reproducible?
4. Are there non-medical studies performed to harmonize images/features? What are the different types of methods?
5. What are the advantages, disadvantages and challenges of various harmonization methods?

Keeping in mind the above research questions, we searched for literature using PubMed and Google Scholar by typing in the following keywords: “radiomics”, “har-monization methods”, “feature reproducibility”, “robustness”, “scanner variation”, “protocol

variation”, “deep learning”, “multicentric studies”, “medical imaging”. Articles were selected based on their novelty, relevance, and being in English.

Image Domain Harmonization

Standardization of Image Acquisition and Reconstruction Parameters across Various Centers Related to Clinical Trials

For multicentric prospective studies, the ideal way to standardize radiomic features is to define and follow imaging protocols that define scanner types in conjunction with acquisition and reconstruction parameters (see Table 1 for summary). For example, the European Society for Therapeutic Radiology and Oncology (ESTRO) panel provides guidelines for procedures and methods for image-guided radiation therapy (IGRT) in prostate cancer [42,43]. This panel consulted a large base of the radiation oncology community from the European Union and developed guidelines for delineating localized prostate cancer in CT and magnetic resonance images (MRI). ESTRO also has a working group focusing on cervical carcinoma for developing and validating methods and imaging parameters from various institutions [44]. For standardization of PET imaging, the European Association of Nuclear Medicine (EANM) [45] launched the EARL (EANM Research Ltd.) program covering areas such as scan acquisition, processing of images, and image interpretation. Pfahler et al. [46] conducted a study to investigate the effects of harmonizing image reconstructions on feature reproducibility and concluded that EARL compliant image reconstruction harmonized a wide selection of radiomic features. A similar initiative by the American Society for Radiation Oncology (ASTRO) [47] was created to develop a ‘practice parameter’, for IGRT and to provide quality assurance standards, personnel qualifications, indications, and guided documentation [48] for imaging. In MRI, however, such guidelines do not exist [49] and most of the MRI modalities are not even quantitative [50]. Efforts have been taken in the past, concerning MRI imaging, for example by UCHealth [51] to reduce the number of MRI protocols from 168 to 66 across scanners and centers by selecting an appropriate clinic-driven protocol and standardization process. Another set of guidelines is provided by the FDA (Food and Drug Administration) [52] to focus on image acquisition in clinical trials conducted to support the authorization of drugs and biological products. Ever since this draft by FDA was released in 2015, it has become a reference standard for most promoters and industries of clinical trials.

Such efforts need to be extended to the radiomics field to help control the variability present across different scanner machines, acquisition and reconstruction parameters. However, these radiomics guidelines might not be able to account for the plethora of existing scanners, protocols, and reconstruction parameters by different vendors across multiple centers.

Table I: Summary table of standardization guidelines/regulations set for image acquisition and reconstruction parameters across various centers.

Standardization of Image Acquisition and Reconstruction Parameters across Various Centers Related to Clinical Trials			
Reference	Data	Variation across	Summary
Mottet et al. ^[42] Cornford et al. ^[43] (ESTRO)	CT and MRI images (prostate cancer)	NA	Provided guidelines procedures and methods for im-age-guided radiation therapy (IGRT) in prostate cancer
Boellaard et al. ^[45] (EARL)	PET imaging	Scan acquisition, image processing, image inter-pretation	Provides guideline/regulations for oncology
Luh et al. ^[48] (ASTRO)	NA	NA	Developed a ‘practice parameter’ for IGRT, and provided quality assurance standards, personnel qualifications, indications and guided documentation for imaging
Sachs et al. ^[51]	CT and MRI images	CT and MRI protocols	Reduced the number of MRI protocols from 168 to 66 and CT protocols from 248 to 97 across scanners and centers by selecting an appropriate clinical-driven protocol and standardization process
Center for Drug Evaluation and Research (FDA) ^[52]	NA	Image acquisition param-eters	Provided guidelines to focus on image acquisition in clini-cal trials conducted to support authorization of drugs and biological products

Post-Processing of Raw Sensor-Level Image Data

It would be worthwhile to work with raw sensor-level data, right before recon-structing the image and apply harmonization methods on it to remove scanner and protocol variability. Image reconstruction, necessary for human viewing and interpre-tation, combined with the manual contouring variability, could lead to a loss of latent raw sensor-level image data and lower precision in measurements. Most machine learning (ML) and DL algorithms have been used on reconstructed images in the existing medical imaging workflow. Instead, the abilities of ML and DL could be leveraged to process the underlying raw sensor-level data to access its hidden nuances [53–55]. A study conducted by Lee et al. [56] investigated the performance of a CNN for classifying raw CT data in the sinogram-space to identify the body region and detect intracranial hemorrhage. The sinogram-specific CNN performed slightly better than the conventional neural network (Inception-V3 [57]) in the image-space by approximately 3% in terms of accuracy. In another study, Gallardo-Estrella et al. [58] proposed a method to reduce variability due to different reconstruction kernels in CT images by decomposing each CT scan into a set of frequency bands and the energy in each frequency band is scaled to a reference value iteratively. This method was validated for emphysema reconstruction. Although this method was applied to normalize fully reconstructed images, the applicability of this method could be extended to harmonize raw image data. Radiomics signature analysis can also be performed directly on the raw image data without the need for reconstruction which adds bias and variability [56,59]. Furthermore, the reconstruction

process itself can also be considered as a prediction problem utilizing raw CT data (sinograms) or k-space values of MRI inputs [60]. These studies widen the scope to apply harmonization methods on raw image data and take advantage of the hidden information in the raw image data rather than applying it in the reconstructed image-space. Refer to Table 2 for a summary of this section.

Table 2: Summary table of post-processing methods of raw sensor-level image data.

Post-Processing of Raw Image Data			
Reference	Data	Variation across	Summary
Lee et al. [56]	Raw sinogram CT data (head and whole-body)	Acquisition parameters in terms of projections and detector like sinograms	Investigated the performance of a CNN for classifying raw CT data in sinogram-space to identify body region and detect intracranial hemorrhage
Gallardo-Estrella et al. [58]	Reconstructed CT images (emphysema in lungs)	Reconstruction kernels	Proposed a method to reduce variability due to different reconstruction kernels in CT images by decomposing each CT scan into a set of frequency bands and the energy in each frequency band is scaled to a reference value iteratively.

Data Augmentation Using GANs

ML-based techniques have emerged to provide effective solutions to translate images across various domains by harmonizing images as opposed to radiomic features alone. Examples include ML-based adaptive dictionary learning [61] and DL methods like using GANs [62–70]. Methods using coefficients of spherical harmonics to harmonize diffusion MRI have been explored [61,71–73]. The applicability of this method was limited to diffusion MRI since the analysis of diffusion MRI requires various processing steps to correct for scanner acquisitions and protocol variation effects and was addressed by the 2018 CDMRI (computational diffusion MRI) Harmonization challenge [74].

Another widely used DL technique in medical image analysis are GANs [75] because of their ability to model target data distributions to generate realistic images (summary in Table 3 at the end of this section). GANs consist of two adversarial networks, a generator that generates realistic data and a discriminator that distinguishes whether the data is real or fake. The objective of a GAN is to keep the generator and discriminator in opposition to each other. Despite the difficulty in handling multi-centric medical data, GANs have shown promising results to overcome the multi-center variation. Zhong et al. [76] used a dual GAN, with U-Net [77] as the backbone, to harmonize the diffusion tensor imaging (DTI) derived metrics on neonatal brains and compared it with three other methods: voxel-wise scaling, global-wise scaling, and ComBat. The results from this study showed that the GAN based method performed better at harmonizing neonatal datasets in multi-centric studies. Another study by Modanwal et al. [78] used a cycleGAN [64] to perform intensity harmonization on MRI breast images obtained from two scanners



(GE and Siemens). A cycleGAN utilizes a cycle consistency loss to translate an image from one domain to another without the requirement for paired data. Cycle consistency loss is an optimization problem in the sense that if a zebra image is converted to a horse image and back to being a zebra image, we should obtain the same input in return. This method was adapted by modifying the discriminator that further helped in preserving the tissue characteristics and shape. This method could operate on unpaired images; however, a downside is that this algorithm worked only for 2D slices and could not retain volume information due to limited computational resources. A comparative study was conducted by Cackowski et al. [79] between ComBat and cycleGAN to harmonize multi-centric MRI images. The authors found that both methods were complementary to each other and had similar effects on the radiomic features. The grey-level run length matrix (GLRLM) features benefited more from ComBat while the cycleGAN performed better on Gray Level Size Zone (GLSZM) features. It would be of great interest to see the effects the combination of ComBat and GAN would have on radiomic features.

Guha et al. [63] conducted a study that transforms low-resolution (LR) CT scans of trabecular (Tb) bone microstructures into high-resolution (HR) CT scans, obtained from two scanners (LR from Siemens FLASH and HR from Siemens FORCE; paired images), using GAN-CIRCLE, of which the architecture is shown in Figure 3. This DL-based method was inspired by You et al. [80] and is monitored by three losses: the identical, residual, and cycle consistency loss. The cycle consistency establishes an end-to-end nonlinear mapping from LR CT to HR CT scans with reference to the Wasserstein distance [81]. This type of loss was first used in cycleGANs [64] and it helps a GAN to perform image-to-image translation between unpaired images by enforcing a strong consistency across domains. The residual network is built to preserve the high frequency anatomical details in the image. The identity loss aids to regularize training by learning sufficient latent structural information to enhance the image resolution. The results were compared to and evaluated against the reference value obtained from the true HR CT scans. The predicted results showed improvement in the structural similarity index with respect to true HR CT scans in terms of Tb network area density, Tb thickness and Tb spacing. Other authors [82–86] have also addressed image up-sampling using DL techniques.

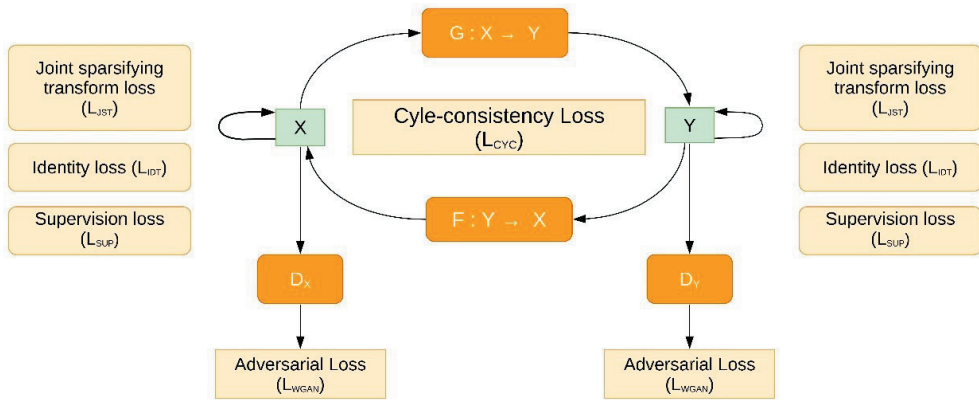


Figure 3: Basic GAN-CIRCLE network. Here, X is a set of LR CT scans and Y is the corresponding HR CT scans. The network has two GAN modules, a low-to-high image reconstructor (Generator G , Dis-criminator D_Y) and a high-to-low image reconstructor (Generator F , Discriminator D_X). Different loss functions are harmoniously coupled for training the network and monitored with regular-ized cycle-consistency and identity loss to prevent overfitting. Figure is adapted from [63].

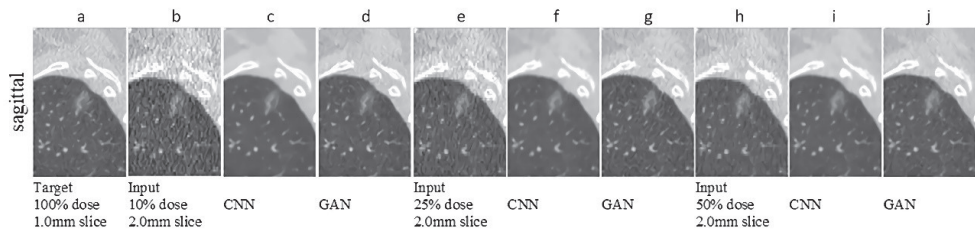


Figure 4: Normalization results reused from [87] with original copy obtained from authors. The row represents a sagittal view of the scans containing the ROI nodule. Column (a) represents the target image; Column (b,e,h) shows the input images with different slice thicknesses and dosage; Col-umns (c,f,i) show the CNN results and Columns (d,g,j) show the GAN-based results.

Style Transfer

The advances in the field of style transfer may prove useful to overcome scanner acquisition and reconstruction parameter variability at the image level. Style transfer is a computer vision technique that requires two images, a content image and a reference style image, and combines them so that the resulting output image preserves the key elements of the content image but appears to be “painted” in the style of the reference style image. When there is no radiomics model available for a new scanner or protocol, style transfer could be applied such that the images coming from a new machine can be transformed so that they look like they were acquired from an existing machine [90]. This section dis-cusses various style transfer methods (Table 4 at the end of this section), starting with the non-CNN methods followed by neural style transfer (NST) methods.

Table 3: Summary table of data augmentation methods using GANs.

Data Augmentation Using GANs			
Reference	Data	Variation across	Summary
Zhong et al. ^[76]	MRI images (neonatal brains)	Scanners, acquisition protocols	Utilized a dual GAN, with U-Net as the backbone to harmonize the diffusion tensor imaging (DTI) derived metrics on neonatal brains
Modanwal et al. ^[78]	MRI images (breast)	Scanners	Utilized a cycleGAN to perform intensity harmonization on MRI breast images obtained from two different scanners
Cackowski et al. ^[79]	MRI images (brain)	Scanners, acquisition protocols	Conducted a comparative study was conducted by Cackowski et al. ^[79] between ComBat and cycleGAN to harmonize multi-centric MRI images
Guha et al. ^[63]	CT images (trabecular bone (Tb) microstructures)	Scanners	Conducted a study that transforms low-resolution (LR) CT scans of trabecular (Tb) bone microstructures into high-resolution (HR) CT scans, obtained from two scanners (LR from Siemens FLASH and HR from Siemens FORCE; paired images), using GAN-CIRCLE ^[80]
Wei et al. ^[87]	CT images (chest)	Dosage, slices thickness	Utilized a 3D GAN to normalize CT images to classify and detect pulmonary nodules

We categorize and briefly explain the existing neural style transfer methods and discuss their strengths and weaknesses.

Before the onset of neural style transfer, image stylization came under the category of non-photorealistic rendering (NPR). Image-based artistic rendering (IB-AR) [91–94] is the artistic stylization of two-dimensional images and can be further categorized into four categories; stroke-based, region-based, example-based, and image processing and filtering. [94]. Stroke-based rendering tries to render strokes (e.g., tiles, stipples or brush strokes) on a content image to adapt to a particular style [95]. However, this method is built to adapt to only one particular style and not arbitrary styles [94]. Region-based rendering [96,97] renders stroke patterns in semantic regions of an image and even though it permits local control over the degree of details, this method also cannot be adapted for arbitrary styles [94]. Hertzmann et al. [98] proposed ‘image analogies’ to learn the mapping between paired source and target images in a supervised fashion but paired images are often not available in practical settings. Even though filtering and image pre-processing [99,100] are efficient and straightforward techniques, they might not be entirely applicable to a wide variety of styles [94]. The above-mentioned techniques do provide dependable stylized results, but their limitations eventually gave rise to novel methods in the field of NST.

The groundbreaking work of Gatys et al. [101] paved the way for a new field of NST. Gatys et al. [101] first conducted a study that separates content from one image and style from another image and combines it into a new image using a neural network (Figure

5). The paper demonstrated that transferring style from one image to the other can be modelled as an optimization problem that can further be solved by training a neural network, VGG-19 [102] in this case. The style was extracted by looking at the spatial correlation between filter responses and this was calculated as the Gram matrix [103] of a feature map. The total loss was calculated as the weighted sum of both content loss (L_c) and style loss (L_s) by weights α and β respectively. Thus, the style transfer task was reduced to creating a new image through an optimization process by minimizing the total loss. However, the high resolution of images affected the speed of the style transfer process and the algorithm failed to preserve the consistency of details and fine structures during style transfer because the low-level information was not retained by the CNN. The Gram matrix is not the only choice for representing style in images. There are also other in-terpretations of Gram matrix, such as MMD mathematically proven by Li et al. [104]. Additionally, the definitions of style and content remain unclear since no representation exists to factorize either style or content of an image.

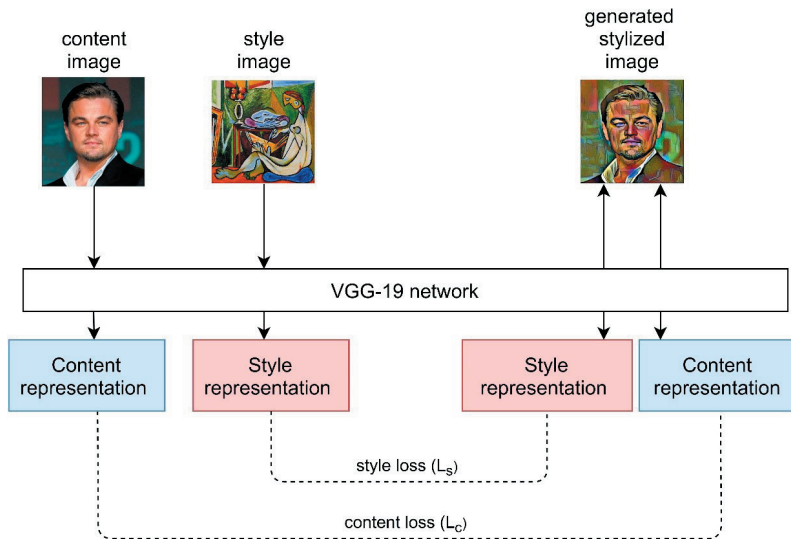


Figure 5: Illustration of concept of neural style transfer using original work [101].

Li et al. [104] questioned the usage of the Gram matrix from Gatys et al. [101] and were not satisfied with the motivation behind its use. They treated neural style transfer as a domain adaptation problem where the difference between the source distribution and the target distribution would be measured and minimized. They provided mathematical proof that matching Gram matrices of filter responses is equivalent to minimizing MMD [105] with the second-order polynomial kernel. The VGG19 network was used here as well, and they proved that the top layers had larger receptive fields and could reproduce more global textures.

Xu et al. [106] proposed a method for arbitrary style transfer, which allows the stylization of images from an unseen content image and style image. They utilized the Behance dataset [107] containing several artistic images and course category labels for style and content. They combined the concepts of original neural style transfer with the concept of adversarial training for arbitrary style transfer from multi-domain images. Xu et al. [106] built a conditional generator to fool the discriminator and to assure that the style and content representations are similar to the input images by combining content and style using adaptive instance normalization (AdaIN) [108]. The method utilized Gram loss for the style representation, perceptual loss [109] for content representation and adversarial loss to capture beyond texture the style information from a distinct style label/category. Their methods outperform previous work using AdaIN [108] and whitening and color transform [110] quantitatively. However, qualitative results in this study show that stylization does not occur beyond a point even after tuning the parameters due to the difficulty of the optimization.

A medically relevant study led by Yang et al. [111] investigated the effects of different kernels on CT images and proposed an unsupervised kernel conversion method by utilizing a cycleGAN with AdaIN [108] that works on unpaired images. They modified the base model of UNet [77] to use polyphase decomposition [112] which resulted in better performance. They assumed that the unsupervised kernel conversion problem can be posed as an unsupervised image style transfer problem that can be solved using optimal transport [113,114]. The qualitative results showed that their methods performed better however, in the quantitative evaluation (peak signal to noise ratio and structure similarity index), supervised learning performed better than unsupervised learning. A similar study by Liu et al. [115] was carried out to harmonize MRI images from multiple arbitrary sites using a style transferable GAN. They treated harmonization as a style transfer problem and proved that their model applied to unseen images provided there was enough data available from multiple sites for training purposes. However, the model only worked on two-dimensional images and not three-dimensional images. They also mention that selecting an appropriate reference image would be challenging if the data pool was vast.

Studies by Armanious et al. [116] and Clancy and Milanko [117] have also utilized the concept of style transfer to perform image-to-image translation between PET-CT images (see Figure 6) and healthy-unhealthy chest X-Rays respectively. The difference between both the studies is that Armanious et al. [116] utilized style transfer losses [101] to match the texture between the stylized image and the target image, while Clancy and Milanko [117] just used the cycleGAN and adversarial losses to perform style transfer. Moreover, the MedGAN created by Armanious et al. [116] incorporates a novel generator CasNet, which is a cascade of UNet blocks to obtain sharper translated images. MedGAN seemed

to outperform other existing image-to-image translation methods (e.g., pix2pix [118] and perceptual adversarial network [119]) by providing quantitative and perceptual assessments. Another study by Fetty et al. [120] investigated how the latent space can be manipulated to obtain high-resolution scans by utilizing their StyleGAN architecture. Their StyleGAN architecture incorporated AdaIN [108] method for transferring style. StyleGAN was trained on MRI to CT images (with pelvic malignancies) and achieved a root mean squared error of 0.34 for CT-MRI translation and a mean absolute error of 59 HU for MRI-CT translation.

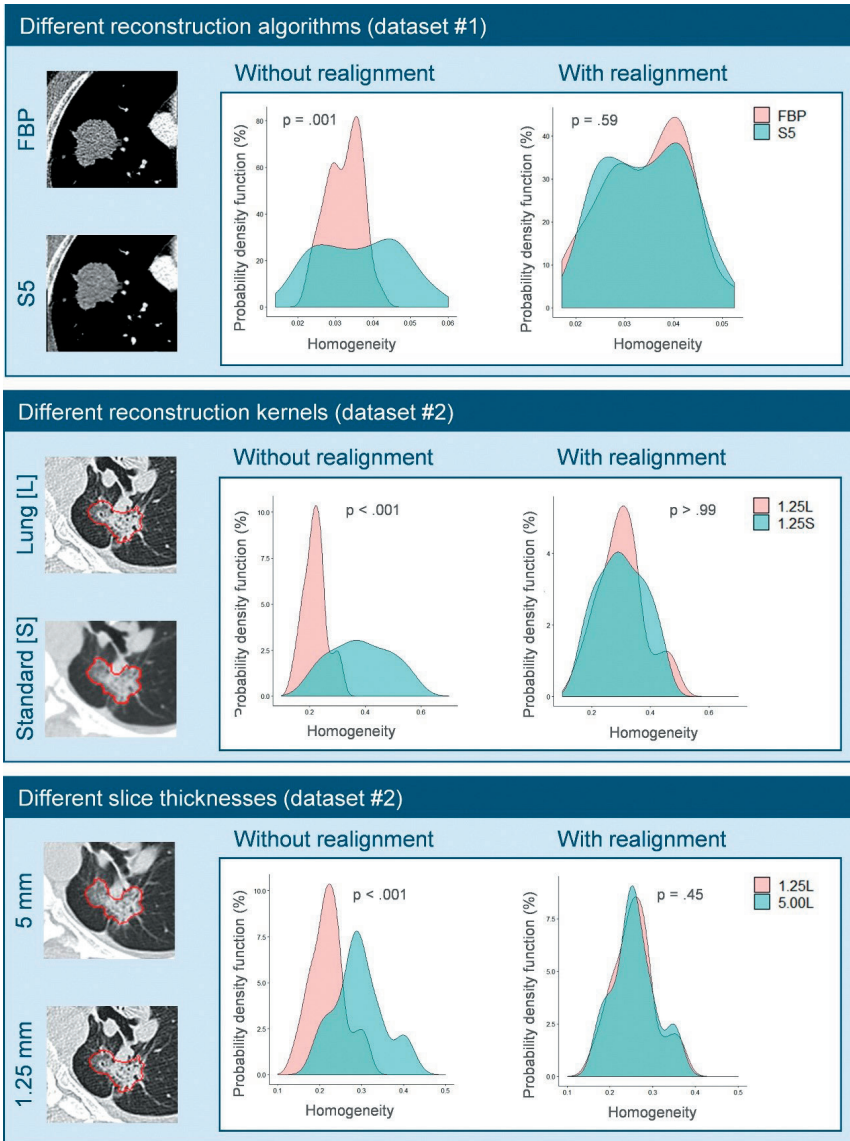


Figure 6: ET to CT translation using MedGAN. Figure reused with under a CC by license from [116].

Many more such studies [121–126] were conducted by applying style transfer methods on medical images, and this approach has the potential to harmonize images, either by image-to-image translations or domain transformations. Depending on the architectures used to perform style transfer, paired or unpaired images might be needed, e.g., if harmonization is to be performed using cycleGAN or StyleGAN as a baseline then paired images are not a requirement. Losses can be modified in such a way that they may or may not include style and content losses from Gatys et al.’s [101] method. In case GANs were to be used for NST, the developer should be mindful of limitations that GANs pose, as discussed in the previous section.

Table 4: Summary table of style transfer methods.

Style Transfer			
Reference	Data	Variation across	Summary
Gatys et al. ^[101]	Non-medical images (mostly artistic images)	NA	Utilized a CNN to perform neural style transfer using Gram matrix
Li et al. ^[104]	Non-medical images (mostly artistic images)	NA	Treated neural style transfer as a domain adaptation problem and proved that matching Gram matrices of filter responses is equivalent to minimizing MMD ^[105]
Xu et al. ^[106]	Non-medical images (mostly artistic images)	NA	Combined the concepts of original neural style transfer with the concept of adversarial training for arbitrary style transfer from multi-domain images
Yang et al. ^[111]	CT images (head, facial bone)	Reconstruction kernels	Investigated the effects of different kernels on CT images and pro-posed an unsupervised image style transfer method by utilizing a cycleGAN with AdaIN ^[108] that works on unpaired images
Liu et al. ^[115]	MRI images (chest)	Multi-center datasets, image acquisition parameters	Harmonized MRI images from multiple arbitrary sites using a style transferable GAN entailing cycle consistency, style and adversarial losses.
Armanious et al. ^[116]	Armanious et al. ^[116]	Multi-modal dataset	Developed MedGAN architecture which consists of a cascade of UNet blocks to obtain sharper translated images (CasNet) along with Gatys et al.’s ^[101] style transfer losses.
Clancy and Milanko ^[117]	X-Rays (chest)	Healthy and unhealthy patients	Utilized the cycleGAN with adversarial losses to perform style transfer.
Fetty et al. ^[120]	MRI, CT images (pelvic malignancies)	Multi-model dataset	Used StyleGAN with baseline GAN architecture and AdaIN method for transferring style across images.

Feature Domain Harmonization

Focusing on Reproducible Features (Identification of Reproducible Features)

These studies test the reproducibility, variability, and repeatability of features extracted from various phantom and patient studies over different reconstruction and acquisition parameters in the case of multi-centric datasets and examine the reproducibility of radiomic features. Refer to Table 5 for a summary.

In the context of PET images, a study by Shiri et al. [127] investigated the impact of various image reconstruction settings on several PET/CT radiomic features obtained from a phantom dataset (developed in-house National Electrical Manufacturers Association [NEMA]) and a patient dataset from two different scanners. Radiomic features were grouped into intensity-based, geometry-based and texture-based features and their reproducibility and variability were evaluated using the coefficient of variation (COV). The results from both phantom and patient studies showed that 47% of all radiomic features were reproducible. Almost half of intensity-based and texture-based and all the geometry-based features were found to be reproducible respectively. The intensity and geometry-based features were also found to be reproducible in another study by Vuong et al. [128], where the authors investigate if the PET/CT radiomics models can be transferred to PET/MRI models by checking the reproducibility of radiomic features against different test-retest and attenuation correction variability. However, Shiri et al. [127] used a phantom body filled with homogeneous activity rather than heterogeneous activity, which does not properly imitate the human tissue. The respiratory motion [127,128], quantization [127,128] and segmentation parameters [127] were also absent in the studies, which may have had a considerable effect on the radiomic features. A similar study by Bailly et al. [129] analyzed the reproducibility of texture features in PET scans across different acquisition and reconstruction parameters in the context of multi-center trials. They found out that only a few features were strongly reproducible and acceptable for multi-center trials. Nevertheless, this study checked the reproducibility of texture features evaluated against reconstruction parameters coming from just one manufacturer. Many such studies have been carried out to check the reproducibility of radiomic features in PET scans [130–144] but most of them only check the impact of variability in scanner and imaging parameters and do not provide concrete image and/or feature harmonization methods to obtain reproducible features.

In the case of CT scans, Prayer et al. [145] conducted a trial to investigate the inter- and intra-scanner repeatability and reproducibility of computed tomography (CT) radiomic features (radiomic feature) of fibrosing interstitial lung disease (fILD). The dataset was obtained from IRB-approved test-retest study with sixty fILD patients. The results showed that intra and inter-scanner reproducibility were highly affected by the variation in slice thicknesses than the variation in reconstruction kernels under study and were reconstruction parameter-specific respectively. The CT radiomic features showed excellent reconstruction parameter-specific repeatability for the test-retest study. However, the sample size of the data used was small, and to check the variability of features only two scanners were used. Careful selection of radiomic features is critical to ensure plausible outcomes in heterogeneous CT datasets. Similar studies have been conducted in the past where the reproducibility of CT radiomic features was investigated using phantom data [25–27,146] as well as patient data [20,147,148]. The phantom

studies were carried forward to reduce the exposure to patients however, they are not real substitutes of heterogeneous human tissues.

Considering MRI, a recent study using radiomics to investigate the reproducibility of features across several MRI scanners and scanning protocol parameters was carried out using both phantom data and patient (volunteer) data by Lee et al. [149]. This study also investigated the repeatability by measuring the variability of radiomic features using a test-retest strategy. The variability of radiomic features across different MRI scanners and protocols was evaluated using the intra-class correlation coefficient (ICC) and the re-peatability was evaluated using the coefficient of variation (COV). The COV measurements showed that there was very little difference in the variability between filtering and normalizing effects which were used for pre-processing. The ICC measurements showed higher repeatability for the phantom data than for the patient data. However, this study was not able to prevent the effects of the volunteer’s movements on the radiomic values despite simulating movements while scanning. A similar study, conducted by Peerlings et al. [150], extracted stable parametric MRI radiomic features with a minimum concordance correlation coefficient of 0.85 between data derived from 61 patients’ test and retest apparent diffusion coefficient (ADC) maps across various MRI-systems, tissues and vendors. A review by Traverso et al. [151] mentions that there are not many phantom studies conducted to investigate the reproducibility of MRI radiomic features. Most of them cover various sites such as the brain [152,153], the gastro-intestinal tract [154–156] and the prostate [157,158], although this limitation was addressed by Rai et al. [159] by developing a novel 3D MRI radiomic phantom to assess the robustness and reproducibility of MRI radiomic features across multiple centers.

Table 5: Summary table of literature which focused on identification of reproducible features.

Focusing on Reproducible Features (Identification of Reproducible Features)			
Reference	Data	Variation across	Summary
Shiri et al. [127]	PET/CT phantom	Image reconstruction settings, scanners	Reproducibility and variability of radiomic features were evaluated using the coefficient of variation (COV)
Bailly et al. [129]	PET scans (gas-troentero-pancreatic neuroendocrine tumors)	Multi-centric trials (acquisition and reconstruction parameters)	Analyzed the reproducibility of textural features in PET scans across different acquisition and reconstruction param-eters in the context of multi-center trials
Prayer et al. [145]	CT scans (fibrosing inter-stitial lung disease (fILD))	Scanners, testretest study	Investigated the inter-and intra-scanner repeatability and reproducibility of computed tomography (CT) radiomic features (radiomic feature) of fILD
Lee et al. [149]	MRI scans (phantom, brain lesions)	Scanners, scanning protocol	Investigated the reproducibility of MRI radiomic features across different MRI scanners and scanning protocol param-eters
Peerlings et al. [150]	MRI scans (ovarian can-cer, colorectal liver me-tastasis)	Vendors, field strengths	Extracted stable parametric MRI radiomic features with a minimum concordance correlation coefficient of 0.85 between data derived from 61 patients’ test and retest apparent diffu-sion coefficient maps

Normalization Techniques

Many statistical normalization methods have been proposed in the past and have calculated the benefits of applying normalization techniques for harmonizing radiomic features affected by variability in scanner acquisition protocols and reconstruction settings.

Statistical Normalization

Chatterjee et al. [160] investigated the effect of applying rescaling and standardization (zero mean, unit standard deviation) as normalization transformations in MRI images obtained from two different institutes with outcome as lymphovascular space invasion and cancer staging. These transformations were applied separately on balanced training and testing sets rather than applying normalization for the entire dataset. This method enhanced the predictive power of the radiomic models through external validation from an external institute. The average prediction accuracy of radiomic features increased from 0.64 to 0.72, average Matthews correlation coefficient (MCC) increased from 0.34 to 0.44 and average F-score increased from 0.48 to 0.71. A similar study by Haga et al. [161] used z-score normalization to standardize the radiomic features extracted from CT images of NSCLC (non-small cell lung cancer) patients from The University of Tokyo Hospital and TCIA (the Cancer Imaging Archive). Z-score normalization uses the formula:

$$z = \frac{(x - \bar{x})}{s} \tag{1}$$

where x is the feature, \bar{x} is the mean and s is the standard deviation and this method gave the best prediction radiomic model with 0.789 AUC (area under the receiver observed characteristics curve) when compared to min-max normalization (0.725 AUC) and whitening from the principle component analysis (0.785 AUC). Refer to Table 6 for a summary.

Table 6: Summary table of basic statistical approaches.

Statistical Normalization			
Reference	Data	Variation across	Summary
Chatterjee et al. [160]	MRI images (endometrial cancer)	Multi-center datasets	Investigated the effect of applying rescaling and standardization as normalization transformations in MRI images obtained from two different institutes. These transformations were applied separately on balanced training and testing sets rather than applying normalization for the entire dataset
Haga et al. [161]	CT images (non-small cell lung cancer (NSCLC))	Multi-centric datasets	Used z-score normalization to standardize the radiomic features extracted from CT images of NSCLC patients from The University of Tokyo Hospital and TCIA (the Cancer Imaging Archive)

Intensity Harmonization Techniques

Crombé et al. [162] performed intensity harmonization techniques (IHT) as a post-processing method on T2-weighted MRI images of sarcoma patients to enhance the MFS (metastatic-relapse-free survival) predictive models. They compared standard normalization, z-score normalization, standardization per signal intensities of healthy tissue, histogram matching and ComBat harmonization methods. A histogram is a statistical representation of an image, which shows the distribution of intensity values. It does not contain information about the location of the image pixels. Histogram matching is where intensity histograms are aligned to a reference intensity histogram. In this study, intensity histogram matching performed better with an AUC of 0.823 in an unsupervised analysis. Related studies [163–165] have used histogram matching to normalize MRI intensity scales. A few studies [166,167] have also applied histogram equalization (enhancing the contrast by flattening the histogram) on images to normalize intensity scales to pre-process images before applying a ComBat harmonization method on top of it. Refer to Table 7 for summary.

Table 7: Summary table of intensity harmonization methods.

Intensity Harmonization Techniques			
Reference	Data	Variation across	Summary
Crombé et al. [162]	MRI images (sarcoma)	Multi-centric datasets	Performed IHT (standard normalization, z-score normalization, standardization per signal intensities of healthy tissue, histogram matching and ComBat harmonization) as a post-processing method on T2-weighted MRI images of sarcoma patients to enhance the MFS (metastatic-relapse-free survival) predictive models
Masson et al. [166]	Contrast enhanced CT images	Multicentric dataset	Applied histogram equalization (enhancing the contrast by flattening the histogram) on images to normalize intensity scales to pre-process images prior to applying ComBat harmonization method on top of it.

ComBat Method and Its Derivatives

ComBat harmonization is a statistical method that was developed originally to harmonize gene expression arrays [168]. ComBat was designed to provide estimates of the effects of assigned batches -which have a single technical difference between each other, while taking into account the effect of biological covariates on the variables or features being harmonized. The estimations are calculated using Bayesian models, and a location/scale shift is performed accordingly to adjust the values of different features. The application of ComBat on radiomics features was first introduced by Fortin et al. [169]. The authors used ComBat to harmonize cortical thickness measurements calculated on diffusion imaging tensor data to remove variations in feature values attributed to differences in acquisition and reconstruction parameters. The authors reported that ComBat removes interscanner variability for these measurements and can also preserve biological correlations. The authors further developed an open software for ComBat that can be used for radiomics analysis.

Following that, several studies further investigated the potential of ComBat harmonization in radiomics analyses. Orlhac et al. [170] investigated the potential of ComBat to correct for the variations of CT radiomic features extracted from scans collected from different centers. The authors reported that all radiomic features were significantly affected by differences in acquisition and reconstruction parameters, and that almost all radiomic features can be used following ComBat harmonization. The authors further reported an improvement in the performance metrics of the developed radiomic signatures after ComBat harmonization. Figure 7 shows the result for this study [170] with three instances of feature distributions realigned between different CT reconstruction algorithms, reconstruction kernels and slice thicknesses. Another study by Orlhac et al. [171] investigated the potential of ComBat to harmonize radiomic features extracted from PET scans acquired differently. The authors reported similar results to that of the application of ComBat on CT scans. A similar study investigated the performance of ComBat harmonization, in addition to modified ComBat methods: M-ComBat, B-ComBat, and BM-ComBat [172]. The study reported a significant improvement in the performance of radiomic signatures following the application of all the investigated ComBat methods.

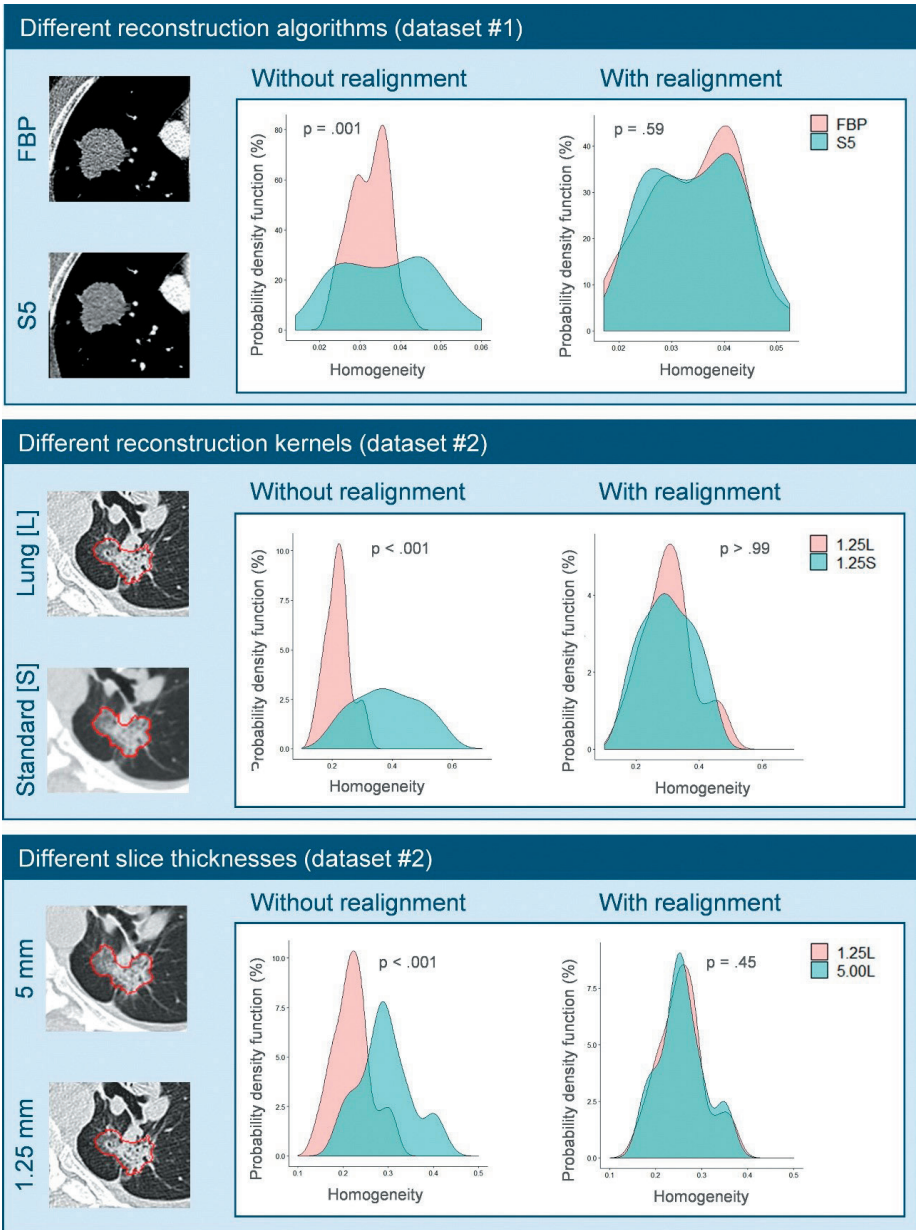


Figure 7: Probability density function of homogeneity before and after applying ComBat for realignment between different CT reconstruction algorithms, reconstruction kernels and slice thicknesses. FBP: filtered back-projection. Figure reproduced from [170]. Figure reproduced with copyright per-mission obtained from The Radiological Society of North America.

Of note, none of the above studies investigated the concordance (reproducibility) of features after ComBat harmonization. Data with similar distributions could still have different individual data points within. Furthermore, the aim of radiomics is to improve personalized medicine. Therefore, for clinical applications, the radiomic signature is expected to be applied on a single patient each time, and not a group of patients simultaneously. Henceforth, the focus of harmonization techniques must be the standardization of radiomic feature values across different imaging settings and patient populations. This is statistically translated into the assessment of concordance in features values following harmonization, and not the performance of developed signatures following harmonization [173].

With regards to the application of ComBat on radiomic features, several points must be taken into consideration: (i) In contrast to gene expression, radiomic features have different complexity levels. Therefore, ComBat is not expected to perform uniformly on all features; (ii) Biological covariates are embedded in the harmonization equation, and as the aim of radiomic studies is to investigate such relationships, biological covariates cannot be provided for the ComBat formula. Furthermore, as the reproducibility of a feature is a cornerstone for it to be further analyzed, solely harmonizing the distribution without paying attention to individual value and rank, is not expected to be beneficial for the generalizability of radiomics signatures. Therefore, the concordance in feature values following ComBat harmonization must be used as an initial feature selection step, to select features that become concordant for further analysis. A framework that guides the use of ComBat in radiomics analyses was published [174]. This framework consists of several steps. The first step is to collect the imaging dataset(s), and to extract the imaging acquisition and reconstruction parameters. Following this, an anthropomorphic phantom is scanned with the different acquisition and reconstruction parameters used for acquiring the scans in the patients' imaging dataset. Radiomic features are then extracted from the phantom scans, and the reproducibility of radiomic features is assessed on those scans using the concordance correlation coefficient (CCC) [175], and the reproducible features ($CCC > 0.9$) could be further used for further modeling. To assess the performance of ComBat, it is applied on the phantom scans, followed by the calculation of the CCC. Radiomic features that obtain a $CCC > 0.9$ following ComBat application are to be considered "ComBatable".

One study applied the framework on thirteen scans of a phantom [176] acquired using different imaging protocols and vendors. The study investigated the reproducibility of radiomic features in a pairwise manner, resulting in a total of seventy-eight pairs. The study reported that different numbers of reproducible radiomic features were identified in each scenario. The results confirmed that radiomic features are affected differently by the differences in imaging protocols and vendors used, with a wide range between nine and seventy-eight reproducible features, substantiating the need for the application of the framework for all radiomic studies [177]. The study also reported that ComBat harmonization did not perform uniformly on radiomic features, and the number of features that could be used following ComBat harmonization ranged between fourteen and eighty radiomic features. Henceforth, the study recommended that the application of ComBat harmonization should follow a similar impact analysis depending on the data under analysis.

Another study utilized a similar framework to assess the performance of ComBat on CT phantom scans that were acquired with the same acquisition and reconstruction parameters except for the in-plane resolution [178,179], on two different scanner models. The authors performed pairwise comparisons between the scans and reported that radiomic features are affected differently by the degree of variation within a single reconstruction parameter (in-plane resolution). A given radiomic feature can be reproducible up to a certain degree of variation in pixel spacing but becomes unreproducible when the variation is relatively large. Other features were found to be reproducible regardless of the variation in pixel spacing, while a few features were found to vary significantly with the slightest change in pixel spacing. These groups of features differed based on the scanner model used to obtain the scans. The application of ComBat on those scans resulted in a different number of reproducible features depending on the variation in the scan in-plane resolution, which also varied according to the scanner model. As such, the study recommended the assessment of the reproducibility and the harmonizability (using any harmonization method) of radiomic features in the data under study before performing radiomics analyses. Refer to Table 8 for summary of ComBat method and its derivatives.

Table 8: Summary table of ComBat methods and its derivatives.

ComBat Method and Its Derivatives			
Reference	Data	Variation across	Summary
Fortin et al. [169]	DTI data	Acquisition and reconstruction parameters	Used ComBat to harmonize cortical thickness measurements calculated on DTI data to remove variations in feature values attributed to differences in acquisition and reconstruction parameters.
Orlhac et al. [170]	CT scans (phantom [180], lung cancer)	Multi-centric dataset	Investigated the potential of ComBat to correct for the variations of CT radiomic features extracted from scans collected from different centers.
Orlhac et al. [171]	PET scans	Acquisition parameters	Investigated the potential of ComBat to harmonize radiomic features extracted from PET scans acquired differently
Ibrahim et al. [174]	Phantom CT [176]	Acquisition and reconstruction parameters	Proposed a framework that guides the use of ComBat in radiomics analyses to assess the performance of ComBat
Ibrahim et al. [177]	Phantom CT [176]	Imaging protocols, vendors	Investigated the reproducibility of radiomic features in a pairwise manner and performed ComBat harmonization on it.
Ibrahim et al. [178,179]	Phantom CT	In-plane resolution	Performed pairwise comparisons between the scans and reported that radiomic features are affected differently by the degree of variation within a single reconstruction parameter (in-plane resolution).

Normalization Using Deep Learning

Andrzejczyk et al. [21] proposed a DL-based technique trained on phantom data to normalize various types of features including hand-crafted and deep features. The main idea is to use a simple neural network (two layers in [21]) to learn a non-linear normalization transformation. This work is based on the assumption that training a deep model on top of features to classify texture types while being adversarial to the scanner of origin creates features that are stable to scanner variations. It therefore aims at reducing in-tra-scan clustering that does not underline true physio-pathological tissue changes, while maintaining highly informative and discriminative features. The generalization of the proposed approach to unseen textures and unseen scanners is demonstrated by a set of experiments using a publicly available CT texture phantom dataset scanned with various imaging devices and parameters. It is assessed by training the model on a subset of classes and scanners and evaluating the stability on the remaining ones. The stability of the normalized features is demonstrated by the increased ICC, clustering based measures showing the class separability, as well as reduced correlation with pixel spacings. The phantom used for this method was developed in [27]. It contains 10 cartridges of different textures and was scanned by 17 different scanners and acquisition settings. Refer to Figure 8 for an overview of their proposed method. Using a phantom allows a controlled analysis that isolates the variation due to scanner variation from other variations related with patient acquisition. Phantoms can also be scanned by specific scanners with special clinical settings to specifically improve the normalization of the features for clinical use. The normalization could therefore be updated to follow

the latest imaging advances and standards. However, while this phantom was designed to mimic actual biomedical tissue types (particularly non-small cell lung cancer), the method has yet to be validated on real patient data.

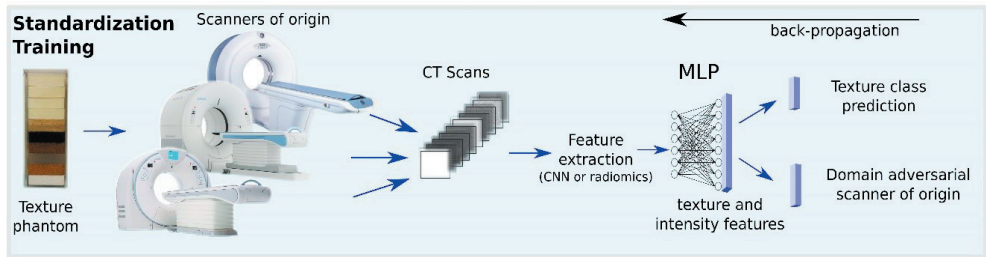


Figure 8: An overview of the proposed normalization method by [21] using a CT phantom. Figure reproduced from [21]. Reproduced from the original copy obtained from authors.

Studies by Rozantsev et al. [181] and Sun and Saenko [182] have adapted divergence-based approaches for domain adaptation by using a two-stream CNN architecture (one in the source domain with synthetic images and the other in the target domain with real images) with unshared weights and the DeepCORAL [183] architecture, respectively. Their methodologies provided a domain-invariant representation by trying to reduce the divergence (reduce the gap/distance) between feature distributions of source and target data distributions (both use non-medical images). Rozantsev et al. [181] used maximum mean discrepancy (MMD) to determine if two samples have the same distribution and Sun and Saenko [182] used correlation alignment that attempts to align the second-order statistics of two distributions by applying a linear transformation. [181] obtains an average accuracy of 0.908 while [182] got an average accuracy of 0.72, both using the Office dataset [184]. However, if these methods were to be applied to medical images, the assumption that scanner information can be eliminated by a simple definable constraint could probably work for linear systems like CT rather than for complex nonlinear systems such as MRI. To make domain adaptation techniques widely applicable, domain adversarial neural networks (DANNs) [185,186] have been explored to increase the invariance of the transformed features to the scanner of origin. DANNs use a label predictor and a domain classifier to optimize the features to make the learned features discriminative for the main task but non-discriminative between the domains. Adapting the same framework as proposed in [186], Dinsdale et al. [187] utilized an iterative update approach that aimed to generate scanner-invariant (i.e., harmonized features) representations of MRI neu-roimages while evaluating the main task (segmentation), thus decreasing the influence of scanner variation on the predictions. Refer to Table 9 for a summary of normalization methods using DL techniques.

Table 9: Summary table of normalization methods using deep learning techniques.

Normalization Using Deep Learning			
Reference	Data	Variation across	Summary
Andrearczyk et al. ^[21]	Phantom CT ^[27]	Acquisition and recon-struction parameters	Proposed a DL-based technique trained on phantom data to normalize various types of features including hand-crafted and deep features using a simple neural network to learn a non-linear normalization transformation
Rozantsev et al. ^[181]	Non-medical images	Synthetic, real image domains	Adapted divergence-based approaches for domain adaptation by using a two-stream CNN architecture (one in the source domain with synthetic images and the other in the target domain with real images) with unshared weights
Sun and Saenko ^[182]	Non-medical images	Different image domain in Office dataset ^[184]	Adapted divergence-based approaches for domain adaptation by using DeepCORAL ^[183] architecture
Dinsdale et al. ^[187]	MRI images (neuro)	Multi-centric dataset	Adapted the framework as in ^[186] and utilized an iterative update approach that aimed to generate scanner-invariant (i.e., harmonized features) representations of MRI neuroimages while evaluating the main task (segmentation).



Discussion

With the emergence of Radiomics within medical image analysis comes the challenges associated with it which could hamper the growth of the field. Both methodologies, traditional hand-crafted features and DL, are faced with standardization issues. The hand-crafted features are most of the time not standardized when the data under analysis is acquired with different scanner acquisition protocols and/or reconstruction settings and there is also a lack of biological correlation of these features. To overcome these limitations, various standardization/harmonization techniques have been introduced and utilized.

In the image domain, the methods mentioned above are applicable on images (raw or reconstructed image). Certain regulations and guidelines can be implemented in imaging protocols by providing quality assurance, indications and guided documentation such as the one laid down by ESTRO and FDA. Such guidelines are not available extensively for MRI [49] but efforts have been taken to reduce the number of MRI protocols [51]. However, these guidelines might not be able to compensate for the number of existing scanners and protocol combinations. Apart from setting guidelines, models can be developed using ML and DL on raw image data or on human interpretable reconstructed images. These methods (data augmentation using GANs/style transfer) have emerged to provide efficient solutions to translate images across various domains to harmonize images rather than the radiomic features. It would be worthwhile to harmonize raw image data with underlying hidden information from the scanners rather than using human interpretable reconstructed images. Studies have been conducted to show that performance of models on raw image data is at par with that of reconstructed images

[56]. Furthermore, GANs have shown promising results to overcome the multi-centric variation. However, GANs are arduous to train due to vanishing gradient challenges that can completely stop the learning process. They are data hungry and suffer from mode collapse causing them to generate similar looking images. On the other extreme, they can also add unrealistic artefacts in the images. Moving forward, advances in the field of style transfer may prove useful to harmonize images without the need of scanner-specific radiomic models. Neural style transfer and its derivatives could extract texture information [101] which could be very useful to obtain reproducible radiomic features in multi-centric trials. Although these techniques have not been specifically analyzed to improve radiomic feature reproducibility, it can be worthwhile to extend their potential to radiomic features.

In the feature domain, various methods have been implemented directly on radiomic features to evaluate its reproducibility and its generalizability across scanner protocol settings. The most convenient and comparatively easy way is to identify reproducible features and focus explicitly on them to evaluate the model's performance. The selection of reproducible features helps build robust models, yet one drawback is that several informative and useful features might be excluded for analyses while extracting 'reproducible features'. Furthermore, there is no generalized threshold for all features above which the latter can be labelled as 'reproducible enough', the condition to be met is that the signal is stronger than the noise. These studies report that variation in scanner acquisition and reconstruction parameters have an impact on the radiomic features and their reproducibility hence highlighting the importance of utilizing harmonization methods for stabilizing radiomic features under analysis. Normalization techniques such as min-max normalization, z-score normalization, histogram matching for intensities, and ComBat harmonization have been explored for radiomic studies. Basic statistical approaches (rescaling/standardization) might be too simplistic to apply considering the fact that some image modalities are complex and non-linear (MRI). Histogram matching or equalization is an efficient method to normalize the intensity scales of images, but it is often used as a pre-processing step to 'clean' the data before feeding it to the radiomics models. On the other hand, ComBat tries to get rid of the 'batch effects' (or scanner/protocol variability) by shifting data distributions while also preserving the biological variation in the data under analysis. However, ComBat relies heavily on labelled data to perform efficient batch correction and estimation [40]. Another disadvantage is that if new data is to be harmonized then it must be added in the existing pool of data for ComBat to perform correctly. Alternatively, normalizing 'deep' features [21] can also be an efficient way to improve the reproducibility of features since DL has a wide scope with various architectures and techniques. Domain adaptation techniques using DL and GANs have the ability to translate images from one domain to another and can thus increase the overlap of feature distribution between two unharmonized images. Data

augmentation, adversarial training, and normalization techniques in combination with neural networks could complement the benefits of neural network training.

Furthermore, to assess the effects that image acquisition parameters have on radiomic features studies have been conducted on phantom images or on images acquired from several different patients to reduce the dosage exposure given to individual patients. One issue with images acquired from different patients is that it introduces high variability due to differences in patient positioning and anatomy [146,188]. On the other hand, objects used for phantom studies are easy to scan for multiple test-retest studies and can be conveniently transported between various imaging sites. Additionally, instructions/guidelines could be set for standardizing the image acquisition parameters to control its variability, tailored to fit the clinical practice. Pre-processing raw sensor-level data is an interesting approach to harmonize images if one wants to make use of the latent information within these raw images. Since a lot of research has already been done using ComBat methods, it would be worthwhile to apply deep learning solutions such as GANs, style transfer, or even normalization using deep learning techniques. These deep learning solutions need further research to show their true potential by applying them to more real medical datasets.

Radiographic phantoms are not the true representatives for realistic patient tissues and this is proved by Mackin et al. [27] who conducted a study showing that the radiomic features extracted from NSCLC (non-small cell lung cancer) and the same features extracted from a phantom (made up of 10 different materials) did not yield the same values for any of the features [27]. Besides, acquisition and reconstruction parameters have also proved to have effects on the radiomic features [20,127,135,189]. Different vendors may have different reconstruction methods and reconstruction parameters that are tailored accordingly at each site/institution.

Conclusions

Radiomics is an emerging field and standardization of radiomic features and/or images is crucial for its survival and impact in this domain when it comes to multicentric studies. Various harmonization methods have been investigated to assess the reproducibility and validation of radiomics across different scanners and protocol settings. This review has covered various topics ranging from methods in the image domain (GANs, style transfer, and regulations guidelines) to methods in the feature domain (statistical normalization, identification of reproducible features, 'deep' feature normalization). The use of harmonization methods has the potential to be beneficial in multi-center studies and the reproducible radiomic features can be practically useful in the decision-making process. Style transfer techniques, with style/content loss or cycle-

consistency loss (e.g., cycleGAN) or in combination, have the potential to harmonize data in the image domain, despite the limitations of GANs. Style transfer needs just two images to work without any prior details about scanners/protocols and hence could be applied on old images in ret-prospective studies and on unpaired images. However, in context of harmonizing images, a limited number of experiments have been conducted and even less for radiomic studies. For harmonization of radiomic features, ComBat methods seem to be extensively used, although normalizing features using deep learning techniques (e.g., domain adaptation methods) can be the way to go ahead too. [21] Showed that normalization using DL can be extended to images coming from unknown scanners and it would be worthwhile to apply this method in combination with GANs in future directions. More work is still needed on identifying limits of features extracted and normalization methods based on just how different the produced images are. Differences linked to scanner model, slice thickness, or reconstruction kernel will likely be in clusters where close clusters can be more easily compared than clusters that are far away from each other. Large datasets of phantom and test-retest data need to be collected for this purpose.

Author Contributions: Conceptualization, S.A.M. and A.C.; methodology, S.A.M. and A.C.; investigation, S.A.M., A.I., H.C.W., V.A. and A.C.; writing—original draft preparation, S.A.M.; writing—review and editing, S.A.M., A.I., H.C.W., V.A., H.M., S.P., Z.S., A.C. and P.L.; supervision, A.C. and P.L.; funding acquisition, P.L. All authors have read and agreed to the published version of the manuscript.

Funding: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812—Hypoximmuno), the European Union's Horizon 2020 research and innovation programme under grant agreement: CHAIMELEON n° 952172, EuCanImage n° 952103, TRANS-CAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295), DRAG-ON—101005122 (H2020-JTI-IMI2-2020-21-single-stage), iCOVID—101016131 (H2020-SC1-PHE-CORONAVIRUS-2020-2) and IMI-OPTIMA n° 10103434.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: Dutch patent filed by A.C. titled 'Method of processing medical images by an analysis system for enabling radiomics signature analysis' Patent no. P127348NL00

References

1. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* 2014, 5, 4006, doi:10.1038/ncomms5006.
2. Hood, L.; Friend, S.H. Predictive, personalized, preventive, participatory (P4) cancer medicine. *Nat. Rev. Clin. Oncol.* 2011, 8, 184–187, doi:10.1038/nrclinonc.2010.227.
3. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* 2012, 48, 441–446, doi:10.1016/j.ejca.2011.11.036.
4. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; et al. Radiomics: The bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 2017, 14, 749–762.
5. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.; Dekker, A.; Fenstermacher, D.; et al. Radiomics: The process and the challenges. *Magn. Reson. Imaging* 2012, 30, 1234–1248, doi:10.1016/j.mri.2012.06.010.
6. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration* 2020, 99, 99–107, doi:10.1159/000505429.
7. Liu, Z.; Wang, S.; Dong, D.; Wei, J.; Fang, C.; Zhou, X.; Sun, K.; Li, L.; Li, B.; Wang, M.; et al. The Applications of Radiomics in Precision Diagnosis and Treatment of Oncology: Opportunities and Challenges. *Theranostics* 2019, 9, 1303–1322, doi:10.7150/thno.30309.
8. Schoolman, H.M.; Bernstein, L.M. Computer use in diagnosis, prognosis, and therapy. *Science* 1978, 200, 926–931, doi:10.1126/science.347580.
9. Zerka, F.; Barakat, S.; Walsh, S.; Bogowicz, M.; Leijenaar, R.T.H.; Jochems, A.; Miraglio, B.; Townend, D.; Lambin, P. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. *JCO Clin. Cancer Inform.* 2020, 4, 184–200, doi:10.1200/cci.19.00047.
10. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016, 278, 563–577, doi:10.1148/radiol.2015151169.
11. Roelofs, E.; Persoon, L.; Nijsten, S.; Wiessler, W.; Dekker, A.; Lambin, P. Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial. *Radiother. Oncol.* 2013, 108, 174–179, doi:10.1016/j.radonc.2012.09.019.
12. Roelofs, E.; Dekker, A.; Meldolesi, E.; van Stiphout, R.G.; Valentini, V.; Lambin, P. International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining. *Radiother. Oncol.* 2014, 110, 370–374, doi:10.1016/j.radonc.2013.11.001.
13. Miotto, R.; Li, L.; Kidd, B.; Dudley, J.T. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Sci. Rep.* 2016, 6, 26094, doi:10.1038/srep26094.

14. Nead, K.T.; Gaskin, G.; Chester, C.; Swisher-McClure, S.; Dudley, J.T.; Leeper, N.J.; Shah, N.H. Androgen Deprivation Therapy and Future Alzheimer's Disease Risk. *J. Clin. Oncol.* 2016, *34*, 566–571, doi:10.1200/jco.2015.63.6266.
15. Gatenby, R.A.; Grove, O.; Gillies, R.J. Quantitative Imaging in Cancer Evolution and Ecology. *Radiology* 2013, *269*, 8–14, doi:10.1148/radiol.13122697.
16. Van Timmeren, J.E.; Cester, D.; Tanadini-Lang, S.; Alkadhhi, H.; Baessler, B. Radiomics in medical imaging—“How-to” guide and critical reflection. *Insights Imaging* 2020, *11*, 1–16, doi:10.1186/s13244-020-00887-2.
17. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* 2021, *8*, 53, doi:10.1186/s40537-021-00444-8.
18. Limkin, E.J.; Sun, R.; Dercle, L.; Zacharaki, E.I.; Robert, C.; Reuzé, S.; Schernberg, A.; Paragios, N.; Deutsch, E.; Ferte, C. Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology. *Ann. Oncol.* 2017, *28*, 1191–1206, doi:10.1093/annonc/mdx034.
19. Vickers, A.J. Prediction Models: Revolutionary in Principle, But Do They Do More Good Than Harm? *J. Clin. Oncol.* 2011, *29*, 2951–2952, doi:10.1200/jco.2011.36.1329.
20. Foy, J.J.; Al-Hallaq, H.; Grekoski, V.; Tran, T.; Guruvadoo, K.; Iii, S.G.A.; Sensakovic, W.F. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: Assessment in a cadaveric liver. *Phys. Med. Biol.* 2020, *65*, 205008, doi:10.1088/1361-6560/abb172.
21. Andrearczyk, V.; Depeursinge, A.; Müller, H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. *J. Med. Imaging* 2019, *6*, 024008, doi:10.1117/1.JMI.6.2.024008.
22. Perone, C.S.; Ballester, P.; Barros, R.; Cohen-Adad, J. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage* 2019, *194*, 1–11, doi:10.1016/j.neuroimage.2019.03.026.
23. Andrearczyk, V.; Oreiller, V.; Jreige, M.; Vallières, M.; Castelli, J.; ElHalawani, H.; Boughdad, S.; Prior, J.O.; Depeursinge, A. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. In *3D Head and Neck Tumor Segmentation in PET/CT Challenge*; Springer: Cham, Switzerland, 2020; pp. 1–21.
24. Zhao, B.; Tan, Y.; Tsai, W.Y.; Schwartz, L.H.; Lu, L. Exploring Variability in CT Characterization of Tumors: A Preliminary Phantom Study. *Transl. Oncol.* 2014, *7*, 88–93, doi:10.1593/tlo.13865.
25. Caramella, C.; Allorant, A.; Orhac, F.; Bidault, F.; Asselain, B.; Ammari, S.; Jaranowski, P.; Moussier, A.; Balleyguier, C.; Lassau, N.; et al. Can we trust the calculation of texture indices of CT images? A phantom study. *Med. Phys.* 2018, *45*, 1529–1536, doi:10.1002/mp.12809.
26. Berenguer, R.; Pastor-Juan, M.D.R.; Canales-Vazquez, J.; Castro-García, M.; Villas, M.V.; Legorbuero, F.M.; Sabater, S. Radi-omics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radiology* 2018, *288*, 407–415, doi:10.1148/

- radiol.2018172361.
27. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; Court, L. Measuring Computed Tomography Scanner Variability of Radiomics Features. *Investig. Radiol.* 2015, 50, 757–765, doi:10.1097/rli.000000000000180.
 28. Cameron, A.; Khalvati, F.; Haider, M.A.; Wong, A. MAPS: A Quantitative Radiomics Approach for Prostate Cancer Detection. *IEEE Trans. Biomed. Eng.* 2016, 63, 1145–1156, doi:10.1109/tbme.2015.2485779.
 29. Parmar, C.; Grossmann, P.; Bussink, J.; Lambin, P.; Aerts, H.J.W.L. Machine Learning methods for Quantitative Radiomic Biomarkers. *Sci. Rep.* 2015, 5, 13087, doi:10.1038/srep13087.
 30. Wibmer, A.; Hricak, H.; Gondo, T.; Matsumoto, K.; Veeraraghavan, H.; Fehr, D.; Zheng, J.; Goldman, D.; Moskowitz, C.; Fine, S.W.; et al. Haralick texture analysis of prostate MRI: Utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different Gleason scores. *Eur. Radiol.* 2015, 25, 2840–2850, doi:10.1007/s00330-015-3701-8.
 31. Peng, Y.; Jiang, Y.; Yang, C.; Brown, J.B.; Antic, T.; Sethi, I.; Schmid-Tannwald, C.; Giger, M.; Eggener, S.E.; Oto, A. Quantitative Analysis of Multiparametric Prostate MR Images: Differentiation between Prostate Cancer and Normal Tissue and Correlation with Gleason Score—A Computer-aided Diagnosis Development Study. *Radiology* 2013, 267, 787–796, doi:10.1148/radiol.13121454.
 32. Cattell, R.; Chen, S.; Huang, C. Robustness of radiomic features in magnetic resonance imaging: Review and a phantom study. *Vis. Comput. Ind. Biomed. Art* 2019, 2, 1–16, doi:10.1186/s42492-019-0025-6.
 33. Jimenez-Del-Toro, O.; Aberle, C.; Bach, M.; Schaer, R.; Obmann, M.M.; Flouris, K.; Konukoglu, E.; Stieltjes, B.; Müller, H.; De-peursinge, A. The Discriminative Power and Stability of Radiomics Features with Computed Tomography Variations: Task-based analysis in an anthropomorphic 3D-printed CT phantom. *Investig. Radiol.* 2021, doi:10.1097/rli.0000000000000795.
 34. Yang, F.; Simpson, G.; Young, L.; Ford, J.; Dogan, N.; Wang, L. Impact of contouring variability on oncological PET radiomics features in the lung. *Sci. Rep.* 2020, 10, 369, doi:10.1038/s41598-019-57171-7.
 35. Pavic, M.; Bogowicz, M.; Würms, X.; Glatz, S.; Finazzi, T.; Riesterer, O.; Roesch, J.; Rudofsky, L.; Friess, M.; Veit-Haibach, P.; et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol.* 2018, 57, 1070–1074, doi:10.1080/0284186x.2018.1445283.
 36. Traverso, A.; Kazmierski, M.; Welch, M.L.; Weiss, J.; Fiset, S.; Foltz, W.D.; Gladwish, A.; Dekker, A.; Jaffray, D.; Wee, L.; et al. Sensitivity of radiomic features to inter-observer variability and image pre-processing in Apparent Diffusion Coefficient (ADC) maps of cervix cancer patients. *Radiother. Oncol.* 2020, 143, 88–94, doi:10.1016/j.radonc.2019.08.008.
 37. Depeursinge, A.; Yanagawa, M.; Leung, A.N.; Rubin, D.L. Predicting adenocarcinoma recurrence using computational texture models of nodule components in lung CT. *Med. Phys.* 2015, 42, 2054–2063, doi:10.1118/1.4916088.

38. Granzier, R.W.Y.; Verbakel, N.M.H.; Ibrahim, A.; Van Timmeren, J.E.; Van Nijnatten, T.J.A.; Leijenaar, R.T.H.; Lobbes, M.B.I.; Smidt, M.L.; Woodruff, H.C. MRI-based radiomics in breast cancer: Feature robustness with respect to inter-observer segmentation variability. *Sci. Rep.* 2020, 10, 14163, doi:10.1038/s41598-020-70940-z.
39. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology* 2020, 295, 328–338, doi:10.1148/radiol.2020191145.
40. Da-Ano, R.; Visvikis, D.; Hatt, M. Harmonization strategies for multicenter radiomics investigations. *Phys. Med. Biol.* 2020, 65, 24TR02, doi:10.1088/1361-6560/aba798.
41. Papadimitroulas, P.; Brocki, L.; Chung, N.C.; Marchadour, W.; Vermet, F.; Gaubert, L.; Eleftheriadis, V.; Plachouris, D.; Visvikis, D.; Kagadis, G.C.; et al. Artificial intelligence: Deep learning in oncological radiomics and challenges of interpretability and data harmonization. *Phys. Med.* 2021, 83, 108–121, doi:10.1016/j.ejmp.2021.03.009.
42. Mottet, N.; Bellmunt, J.; Bolla, M.; Briers, E.; Cumberbatch, M.G.; De Santis, M.; Fossati, N.; Gross, T.; Henry, A.M.; Joniau, S.; et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part 1: Screening, Diagnosis, and Local Treatment with Curative Intent. *Eur. Urol.* 2017, 71, 618–629, doi:10.1016/j.eururo.2016.08.003.
43. Cornford, P.; Bellmunt, J.; Bolla, M.; Briers, E.; De Santis, M.; Gross, T.; Henry, A.M.; Joniau, S.; Lam, T.B.; Mason, M.D.; et al. EAU-ESTRO-SIOG Guidelines on Prostate Cancer. Part II: Treatment of Relapsing, Metastatic, and Castration-Resistant Prostate Cancer. *Eur. Urol.* 2017, 71, 630–642, doi:10.1016/j.eururo.2016.08.002.
44. Pötter, R.; Haie-Meder, C.; Van Limbergen, E.; Barillot, I.; De Brabandere, M.; Dimopoulos, J.; Dumas, I.; Erickson, B.; Lang, S.; Nulens, A.; et al. Recommendations from gynaecological (GYN) GEC ESTRO working group (II): Concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy—3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology. *Radiother. Oncol.* 2006, 78, 67–77, doi:10.1016/j.radonc.2005.11.014.
45. Boellaard, R.; Delgado-Bolton, R.; Oyen, W.J.G.; Giammarile, F.; Tatsch, K.; Eschner, W.; Verzijlbergen, F.J.; Barrington, S.F.; Pike, L.C.; Weber, W.A.; et al. FDG PET/CT: EANM procedure guidelines for tumour imaging: Version 2.0. *Eur. J. Nucl. Med. Mol. Imaging* 2015, 42, 328–354, doi:10.1007/s00259-014-2961-x.
46. Pfaehler, E.; Van Sluis, J.; Merema, B.B.; van Ooijen, P.; Berendsen, R.C.; Van Velden, F.H.; Boellaard, R. Experimental Multi-center and Multivendor Evaluation of the Performance of PET Radiomic Features Using 3-Dimensionally Printed Phantom Inserts. *J. Nucl. Med.* 2020, 61, 469–476, doi:10.2967/jnumed.119.229724.
47. Yu JB, Beck TF, Anscher MS, et al (2019) Analysis of the 2017 American Society for Radiation Oncology (ASTRO) Research Portfolio. *Int J Radiat Oncol Biol Phys* 103:297–304. <https://doi.org/10.1016/j.ijrobp.2018.07.2056>
48. Luh, J.Y.; Albuquerque, K.V.; Cheng, C.; Ermoian, R.P.; Nabavizadeh, N.; Parsai, H.; Roeske, J.C.;

- Weiss, S.E.; Wynn, R.B.; Yu, Y.; et al. ACR–ASTRO Practice Parameter for Image-guided Radiation Therapy (IGRT). *Am. J. Clin. Oncol.* 2020, 43, 459–468, doi:10.1097/coc.0000000000000697.
49. Carré, A.; Klausner, G.; Edjlali, M.; Lerousseau, M.; Briend-Diop, J.; Sun, R.; Ammari, S.; Reuzé, S.; Andres, E.A.; Estienne, T.; et al. Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics. *Sci. Rep.* 2020, 10, 12340, doi:10.1038/s41598-020-69298-z.
 50. Schick, U.; Lucia, F.; Dissaux, G.; Visvikis, D.; Badic, B.; Masson, I.; Pradier, O.; Bourbonne, V.; Hatt, M. MRI-derived radi-omics: Methodology and clinical applications in the field of pelvic oncology. *Br. J. Radiol.* 2019, 92, 20190105, doi:10.1259/bjr.20190105.
 51. Sachs, P.B.; Hunt, K.; Mansoubi, F.; Borgstede, J. CT and MR Protocol Standardization across a Large Health System: Providing a Consistent Radiologist, Patient, and Referring Provider Experience. *J. Digit. Imaging* 2016, 30, 11–16, doi:10.1007/s10278-016-9895-8.
 52. Center for Drug Evaluation and Research. Clinical Trial Imaging Endpoint Process Standards. 2020. Available online: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-trial-imaging-endpoint-process-standards-guidance-industry> (accessed on 12 April 2021).
 53. Chung, C.; Kalpathy-Cramer, J.; Knopp, M.V.; Jaffray, D.A. In the Era of Deep Learning, Why Reconstruct an Image at All? *J. Am. Coll. Radiol.* 2021, 18, 170–173, doi:10.1016/j.jacr.2020.09.050.
 54. De Man, Q.; Haneda, E.; Claus, B.; Fitzgerald, P.; De Man, B.; Qian, G.; Shan, H.; Min, J.; Sabuncu, M.; Wang, G. A two-dimensional feasibility study of deep learning-based feature detection and characterization directly from CT sinograms. *Med. Phys.* 2019, 46, e790–e800, doi:10.1002/mp.13640.
 55. Gao, Y.; Tan, J.; Liang, Z.; Li, L.; Huo, Y. Improved computer-aided detection of pulmonary nodules via deep learning in the sinogram domain. *Vis. Comput. Ind. Biomed. Art* 2019, 2, 1–9, doi:10.1186/s42492-019-0029-2.
 56. Lee, H.; Huang, C.; Yune, S.; Tajmir, S.H.; Kim, M.; Do, S. Machine Friendly Machine Learning: Interpretation of Computed Tomography without Image Reconstruction. *Sci. Rep.* 2019, 9, 15540, doi:10.1038/s41598-019-51779-5.
 57. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016*; pp. 2818–2826, doi:10.1109/cvpr.2016.308.
 58. Gallardo-Estrella, L.; Lynch, D.A.; Prokop, M.; Stinson, D.; Zach, J.; Judy, P.F.; Van Ginneken, B.; Van Rikxoort, E.M. Normal-izing computed tomography data reconstructed with different filter kernels: Effect on emphysema quantification. *Eur. Radiol.* 2016, 26, 478–486, doi:10.1007/s00330-015-3824-y.
 59. Lambin, P.; Woodruff, H. Method of Performing Radiomics Analysis on Image Data. Application No.: N2028271; Ref: P129643NL01, 21 May 2021.
 60. Ravishankar, S.; Ye, J.C.; Fessler, J.A. Image Reconstruction: From Sparsity to Data-Adaptive Methods and Machine Learning. *Proc. IEEE Inst. Electr. Electron. Eng.* 2019, 108, 86–109,

- doi:10.1109/jproc.2019.2936204.
61. St-Jean, S.; Viergever, M.A.; Leemans, A. Harmonization of diffusion MRI datasets with adaptive dictionary learning. *Hum. Brain Mapp.* 2020, 41, 4478–4499.
 62. Dewey, B.E.; Zhao, C.; Reinhold, J.C.; Carass, A.; Fitzgerald, K.C.; Sotirchos, E.S.; Saidha, S.; Oh, J.; Pham, D.L.; Calabresi, P.A.; et al. DeepHarmony: A deep learning approach to contrast harmonization across scanner changes. *Magn. Reson. Imaging* 2019, 64, 160–170, doi:10.1016/j.mri.2019.05.041.
 63. Guha, I.; Nadeem, S.A.; You, C.; Zhang, X.; Levy, S.M.; Wang, G.; Torner, J.C.; Saha, P.K. Deep learning based high-resolution reconstruction of trabecular bone microstructures from low-resolution CT scans using GAN-CIRCLE. *Proc. SPIE Int. Soc. Opt. Eng.* 2020, 11317, 113170U, doi:10.1117/12.2549318.
 64. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Net-works. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017*.
 65. Zhao, F.; the UNC/UMN Baby Connectome Project Consortium; Wu, Z.; Wang, L.; Lin, W.; Xia, S.; Shen, D.; Li, G. Harmoniza-tion of Infant Cortical Thickness Using Surface-to-Surface Cycle-Consistent Adversarial Networks. *Med. Image Comput. Comput. Assist. Interv.* 2019, 11767, 475–483, doi:10.1007/978-3-030-32251-9_52.
 66. Kim, T.; Cha, M.; Kim, H.; Lee, J.K.; Kim, J. Learning to Discover Cross-Domain Relations with Generative Adversarial Net-works. *arXiv* 2017, arXiv:1703.05192.
 67. Yi, Z.; Zhang, H.; Tan, P.; Gong, M. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision, Cambridge, MA, USA, 20–23 June 2017*; pp. 2849–2857.
 68. Goodfellow, I. NIPS 2016 Tutorial: Generative Adversarial Networks. *arXiv* 2016, arXiv:1701.00160.
 69. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; et al. *Generative Adversarial Nets*. In *Mining of Massive Datasets*; Cambridge University Press: Cambridge, UK, 2014.
 70. Yan, C.; Lin, J.; Li, H.; Xu, J.; Zhang, T.; Chen, H.; Woodruff, H.C.; Wu, G.; Zhang, S.; Xu, Y.; et al. Cycle-Consistent Generative Adversarial Network: Effect on Radiation Dose Reduction and Image Quality Improvement in Ultralow-Dose CT for Eval-uation of Pulmonary Tuberculosis. *Korean J. Radiol.* 2021, 22, 983–993, doi:10.3348/kjr.2020.0988.
 71. Moyer, D.; Steeg, G.V.; Tax, C.M.W.; Thompson, P.M. Scanner invariant representations for diffusion MRI harmonization. *Magn. Reson. Med.* 2020, 84, 2174–2189, doi:10.1002/mrm.28243.
 72. Mirzaalian, H.; de Pierrefeu, A.; Savadjiev, P.; et al. Harmonizing Diffusion MRI Data across Multiple Sites and Scanners. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 12–19.
 73. Mirzaalian, H.; Ning, L.; Savadjiev, P.; Pasternak, O.; Bouix, S.; Michailovich, O.; Grant, G.; Marx, C.; Morey, R.; Flashman, L.; et al. Inter-site and inter-scanner diffusion MRI data harmonization. *NeuroImage* 2016, 135, 311–323, doi:10.1016/j.neuroimage.2016.04.041.

74. Tax, C.M.; Grussu, F.; Kaden, E.; Ning, L.; Rudrapatna, U.; Evans, C.J.; St-Jean, S.; Leemans, A.; Koppers, S.; Merhof, D.; et al. Cross-scanner and cross-protocol diffusion MRI data harmonisation: A benchmark database and evaluation of algorithms. *NeuroImage* 2019, 195, 285–299, doi:10.1016/j.neuroimage.2019.01.077.
75. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.C.; Bengio, Y. Generative Ad-versarial Networks. *ArXiv*, abs/1406.2661.
76. Zhong, J.; Wang, Y.; Li, J.; Xue, X.; Liu, S.; Wang, M.; Gao, X.; Wang, Q.; Yang, J.; Li, X. Inter-site harmonization based on dual generative adversarial networks for diffusion tensor imaging: Application to neonatal white matter development. *Biomed. Eng. Online* 2020, 19, 1–18, doi:10.1186/s12938-020-0748-9.
77. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, 16–20 Septem-ber; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241, doi:10.1007/978-3-319-24574-4_28.
78. Modanwal, G.; Vellal, A.; Buda, M.; Mazurowski, M.A. MRI image harmonization using cycle-consistent generative adver-sarial network. In *Medical Imaging 2020: Computer-Aided Diagnosis*; International Society for Optics and Photonics: Belling-ham, WA, USA, 2020; p. 1131413.
79. Cackowski, S.; Barbier, E.L.; Dojat, M.; Christen, T. comBat versus cycleGAN for multi-center MR images harmonization. *Proc. Mach. Learn. Res.* 2021, 1–15.
80. You, C.; Cong, W.; Vannier, M.W.; Saha, P.K.; Hoffman, E.; Wang, G.; Li, G.; Zhang, Y.; Zhang, X.; Shan, H.; et al. CT Su-per-Resolution GAN Constrained by the Identical, Residual, and Cycle Learning Ensemble (GAN-CIRCLE). *IEEE Trans. Med. Imaging* 2020, 39, 188–203, doi:10.1109/tmi.2019.2922960.
81. Arjovsky, M.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* 2017, arXiv:1701.07875.
82. Chen, Y.; Shi, F.; Christodoulou, A.G.; Xie, Y.; Zhou, Z.; Li, D. Efficient and Accurate MRI Super-Resolution Using a Generative Adversarial Network and 3D Multi-level Densely Connected Network. In *International Conference on Medical Image Compu-ting and Computer-Assisted Intervention*; Metzler, J.B., Ed.; Springer: Cham, Switzerland, 2018; pp. 91–99.
83. Park, J.; Hwang, D.; Kim, K.Y.; Kang, S.K.; Kim, Y.K.; Lee, J.S. Computed tomography super-resolution using deep convolu-tional neural network. *Phys. Med. Biol.* 2018, 63, 145011, doi:10.1088/1361-6560/aacdd4.
84. Yu, H.; Liu, D.; Shi, H.; Wang, Z.; Wang, X.; Cross, B.; Bramler, M.; Huang, T.S. Computed tomography super-resolution using convolutional neural networks. In *Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP)*, Beijing, China, 17–20 September 2017; pp. 3944–3948.
85. Chaudhari, A.S.; Fang, Z.; Kogan, F.; Wood, J.; Stevens, K.J.; Gibbons, E.K.; Lee, J.H.; Gold, G.E.; Hargreaves, B.A. Super-resolution musculoskeletal MRI using deep learning. *Magn. Reson. Med.* 2018, 80, 2139–2154, doi:10.1002/mrm.27178.
86. Wolterink, J.M.; Leiner, T.; Viergever, M.A.; Isgum, I. Generative Adversarial Networks for Noise Reduction in Low-Dose CT. *IEEE Trans. Med. Imaging* 2017, 36, 2536–2545, doi:10.1109/

- tmi.2017.2708987.
87. Wei, L.; Lin, Y.; Hsu, W. Using a Generative Adversarial Network for CT Normalization and Its Impact on Radiomic Features. In Proceedings of the 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), Iowa City, IA, USA, 3–7 April 2020; pp. 844–848.
 88. Miyato, T.; Kataoka, T.; Koyama, M.; Yoshida, Y. Spectral Normalization for Generative Adversarial Networks. arXiv 2018, arXiv:1802.05957.
 89. Castiglioni, I.; Rundo, L.; Codari, M.; Di Leo, G.; Salvatore, C.; Interlenghi, M.; Gallivanone, F.; Cozzi, A.; D’Amico, N.C.; Sar-danelli, F. AI applications to medical images: From machine learning to deep learning. *Phys. Med.* 2021, 83, 9–24, doi:10.1016/j.ejmp.2021.02.006.
 90. Chatterjee, A. Method of Processing Medical Images by an Analysis System for Enabling Radiomics Signature Analysis. Patent 2020 No. P127348NL00.
 91. Rosin, P.; Collomosse, J. *Image and Video-Based Artistic Stylisation*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2012.
 92. Kyprianidis, J.E.; Collomosse, J.; Wang, T.; Isenberg, T. State of the “Art”: A Taxonomy of Artistic Stylization Techniques for Images and Video. *IEEE Trans. Vis. Comput. Graph.* 2012, 19, 866–885, doi:10.1109/tvcg.2012.160.
 93. Semmo, A.; Isenberg, T.; Döllner, J. Neural style transfer: A paradigm shift for image-based artistic rendering? In Proceedings of the Symposium on Non-Photorealistic Animation and Rendering, Los Angeles, CA, USA, 29–30 July 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 1–13.
 94. Jing, Y.; Yang, Y.; Feng, Z.; Ye, J.; Yu, Y.; Song, M. Neural Style Transfer: A Review. *IEEE Trans. Vis. Comput. Graph.* 2020, 26, 3365–3385, doi:10.1109/tvcg.2019.2921336.
 95. Hertzmann, A. Painterly rendering with curved brush strokes of multiple sizes. In Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, Orlando, FL, USA, 19–24 July 1998; pp. 453–460.
 96. Kolliopoulos, A. Image Segmentation for Stylized Non-Photorealistic Rendering and Animation. Available online: <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.3711&rep=rep1&type=pdf> (accessed on 24 March 2021).
 97. Gooch, B.; Coombe, G.; Shirley, P. Artistic Vision: Painterly rendering using computer vision techniques. In Proceedings of the 2nd International Symposium on Non-Photorealistic Animation and Rendering, Annecy, France, 3–5 June 2002; Association for Computing Machinery: New York, NY, USA, 2002; p. 83.
 98. Hertzmann, A.; Jacobs, C.E.; Oliver, N.; Curless, B.; Salesin, D.H. Image analogies. In Proceedings of the 28th annual conference on Computer graphics and interactive techniques—SIGGRAPH ’01, Los Angeles, CA, USA, 28 July–1 August 2001; pp. 327–340.
 99. Winnemöller, H.; Olsen, S.C.; Gooch, B. Real-time video abstraction. *ACM Trans. Graph.* 2006, 25, 1221–1226, doi:10.1145/1141911.1142018.
 100. Gooch, B.; Reinhard, E.; Gooch, A. Human facial illustrations: Creation and psychophysical evaluation. *ACM Trans. Graph.* 2004, 23, 27–44, doi:10.1145/966131.966133.
 101. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image Style Transfer Using Convolutional Neural Networks.

- In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2414–2423.
102. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.
 103. Gatys, L.A.; Ecker, A.S.; Bethge, M. Texture synthesis using convolutional neural networks. arXiv 2015, arXiv:1505.07376.
 104. Li, Y.; Wang, N.; Liu, J.; Hou, X. Demystifying Neural Style Transfer. In Proceedings of the Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, Melbourne, Australia, 19–25 August 2017; pp. 2230–2236.
 105. Gretton, A.; Borgwardt, K.M.; Rasch, M.J. A kernel two-sample test. *J. Mach. Learn.* 2012, 13, 723–773.
 106. Xu, Z.; Wilber, M.J.; Fang, C.; Hertzmann, A.; Jin, H. Learning from Multi-domain Artistic Images for Arbitrary Style Transfer. arXiv 2018, arXiv:1805.09987.
 107. Wilber, M.J.; Fang, C.; Jin, H.; Hertzmann, A.; Collomosse, J.; Belongie, S. BAM! The Behance Artistic Media Dataset for Recognition beyond Photography. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1211–1220.
 108. Huang, X.; Belongie, S. Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1501–1510.
 109. Johnson, J.; Alahi, A.; Fei-Fei, L. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *Computer Vision—ECCV 2016*; Springer: Cham, Switzerland, 2016; pp. 694–711, doi:10.1007/978-3-319-46475-6_43.
 110. Li, Y.; Fang, C.; Yang, J.; Wang, Z.; Lu, X.; Yang, M. Universal Style Transfer via Feature Transforms. arXiv 2017, arXiv:1705.08086.
 111. Yang, S.; Kim, E.Y.; Ye, J.C. Continuous Conversion of CT Kernel using Switchable CycleGAN with AdaIN. arXiv 2020, arXiv:2011.13150.
 112. Kim, B.; Ye, J.C. Mumford–Shah Loss Functional for Image Segmentation with Deep Learning. *IEEE Trans. Image Process.* 2020, 29, 1856–1866, doi:10.1109/tip.2019.2941265.
 113. Villani, C. *Optimal Transport: Old and New*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2008.
 114. Peyré, G.; Cuturi, M. Computational Optimal Transport: With Applications to Data Science. *Found. Trends Mach. Learn.* 2019, 11, 355–607, doi:10.1561/22000000073.
 115. Liu, M.; Maiti, P.; Thomopoulos, S.I.; Zhu, A.; Chai, Y.; Kim, H.; Jahanshad, N. Style Transfer Using Generative Adversarial Networks for Multi-Site MRI Harmonization. *bioRxiv* 2021, doi:10.1101/2021.03.17.435892.
 116. Armanious, K.; Jiang, C.; Fischer, M.; et al. MedGAN: Medical Image Translation using GANs. arXiv 2018, arXiv:1806.06397.
 117. Clancy, T.; Milanko, B. Applications of Cyclic Invariant Style Transfers in Medical Imaging. 2019. URL: https://benmilanko.com/projects/chestgan/style_transfers_in_medical_imaging.pdf
 118. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial

- Networks. arXiv 2016, arXiv:1611.07004.
119. Wang, C.; Xu, C.; Wang, C.; Tao, D. Perceptual Adversarial Networks for Image-to-Image Transformation. *IEEE Trans. Image Process.* 2018, 27, 4066–4079, doi:10.1109/tip.2018.2836316.
 120. Fetty, L.; Bylund, M.; Kuess, P.; Heilemann, G.; Nyholm, T.; Georg, D.; Löfstedt, T. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. *Z. Med. Phys.* 2020, 30, 305–314, doi:10.1016/j.zemedi.2020.05.001.
 121. Ma, C.; Ji, Z.; Gao, M. Neural Style Transfer Improves 3D Cardiovascular MR Image Segmentation on Inconsistent Data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2019; pp. 128–136.
 122. Xu, Y.; Li, Y.; Shin, B.-S. Medical image processing with contextual style transfer. *Hum. Cent. Comput. Inf. Sci.* 2020, 10, 1–16, doi:10.1186/s13673-020-00251-9.
 123. Nishar, H.; Chavanke, N.; Singhal, N. Histopathological Stain Transfer Using Style Transfer Network with Adversarial Loss. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2020; pp. 330–340.
 124. Ganesh, A.; Vasanth, N.R.; George, K. Staining of Histopathology Slides Using Image Style Transfer Algorithm. In *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, Bengaluru, India, 18–21 November 2018; pp. 1254–1260.
 125. Nyíri, T.; Kiss, A. Style Transfer for Dermatological Data Augmentation. In *Proceedings of SAI Intelligent Systems Conference*; Springer: Cham, Switzerland, 2020; pp. 915–923.
 126. Yamashita, R.; Long, J.; Banda, S.; et al. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. arXiv 2021, arXiv:2102.01678.
 127. Shiri, I.; Rahmim, A.; Ghafarian, P.; Geramifar, P.; Abdollahi, H.; Bitarafan-Rajabi, A. The impact of image reconstruction settings on 18F-FDG PET radiomic features: Multi-scanner phantom and patient studies. *Eur. Radiol.* 2017, 27, 4498–4509, doi:10.1007/s00330-017-4859-z.
 128. Vuong, D.; Tanadini-Lang, S.; Huellner, M.W.; Veit-Haibach, P.; Unkelbach, J.; Andratschke, N.; Kraft, J.; Guckenberger, M.; Bogowicz, M. Interchangeability of radiomic features between [18F]-FDG PET/CT and [18F]-FDG PET/MR. *Med. Phys.* 2019, 46, 1677–1685.
 129. Bailly, C.; Bodet-Milin, C.; Couespel, S.; Necib, H.; Kraeber-Bodéré, F.; Ansquer, C.; Carlier, T. Revisiting the Robustness of PET-Based Textural Features in the Context of Multi-Centric Trials. *PLoS ONE* 2016, 11, e0159984, doi:10.1371/journal.pone.0159984.
 130. Pfahler, E.; Beukinga, R.J.; De Jong, J.R.; Slart, R.H.J.A.; Slump, C.H.; Dierckx, R.A.J.O.; Boellaard, R. Repeatability of 18 F- FDG PET radiomic features: A phantom study to explore sensitivity to image reconstruction settings, noise, and delineation method. *Med. Phys.* 2019, 46, 665–678, doi:10.1002/mp.13322.
 131. Van Timmeren, J.E.; Carvalho, S.; Leijenaar, R.T.H.; Troost, E.G.C.; Van Elmpt, W.; De Ruyscher, D.; Muratet, J.-P.; Denis, F.; Schimek-Jasch, T.; Nestle, U.; et al. Challenges and caveats of a multi-center retrospective radiomics study: An example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics. *PLoS ONE* 2019, 14, e0217536, doi:10.1371/

- journal.pone.0217536.
132. Orlhac, F.; Soussan, M.; Maisonobe, J.-A.; Garcia, C.A.; Vanderlinden, B.; Buvat, I. Tumor Texture Analysis in 18F-FDG PET: Relationships Between Texture Parameters, Histogram Indices, Standardized Uptake Values, Metabolic Volumes, and Total Lesion Glycolysis. *J. Nucl. Med.* 2014, 55, 414–422, doi:10.2967/jnumed.113.129858.
 133. Cortes-Rodicio, J.; Sanchez-Merino, G.; Garcia-Fidalgo, M.; Tobalina-Larrea, I. Identification of low variability textural features for heterogeneity quantification of 18F-FDG PET/CT imaging. *Rev. Española Med. Nucl. Imagen Mol.* 2016, 35, 379–384, doi:10.1016/j.remnm.2016.04.002.
 134. Nyflot, M.J.; Yang, F.; Byrd, D.; Bowen, S.R.; Sandison, G.A.; Kinahan, P. Quantitative radiomics: Impact of stochastic effects on textural feature analysis implies the need for standards. *J. Med. Imaging* 2015, 2, 041002, doi:10.1117/1.jmi.2.4.041002.
 135. Galavis, P.; Hollensen, C.; Jallow, N.; Paliwal, B.; Jeraj, R. Variability of textural features in FDG PET images due to different acquisition modes and reconstruction parameters. *Acta Oncol.* 2010, 49, 1012–1016, doi:10.3109/0284186x.2010.498437.
 136. Van Velden, F.H.P.; Kramer, G.M.; Frings, V.; et al. Repeatability of Radiomic Features in Non-Small-Cell Lung Cancer [18F]FDG-PET/CT Studies: Impact of Reconstruction and Delineation. *Mol. Imaging Biol.* 2016, 18, 788–795.
 137. Leijenaar, R.T.; Nalbantov, G.; Carvalho, S.; Van Elmpt, W.J.; Troost, E.G.; Boellaard, R.; Aerts, H.J.; Gillies, R.J.; Lambin, P. The effect of SUV discretization in quantitative FDG-PET Radiomics: The need for standardized methodology in tumor texture analysis. *Sci. Rep.* 2015, 5, 11075, doi:10.1038/srep11075.
 138. Shiri, I.; Rahmim, A.; Abdollahi, H.; et al. Radiomics texture features variability and reproducibility in advance image re-construction setting of oncological PET/CT. *Eur. J. Nucl. Med. Mol. Imaging* 2016, 43, S150–S150.
 139. Yan, J.; Chu-Shern, J.L.; Loi, H.Y.; Khor, L.K.; Sinha, A.K.; Quek, S.T.; Tham, I.W.; Townsend, D.W. Impact of Image Reconstruction Settings on Texture Features in 18F-FDG PET. *J. Nucl. Med.* 2015, 56, 1667–1673, doi:10.2967/jnumed.115.156927.
 140. Belli, M.L.; Mori, M.; Broggi, S.; Cattaneo, G.M.; Bettinardi, V.; Dell’Oca, I.; Fallanca, F.; Passoni, P.; Vanoli, E.G.; Calandrino, R.; et al. Quantifying the robustness of [18 F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Phys. Med.* 2018, 49, 105–111, doi:10.1016/j.ejmp.2018.05.013.
 141. Desseroit, M.C.; Tixier, F.; Weber, W.A. Reliability of PET/CT shape and heterogeneity features in functional and morphologic components of non-small cell lung cancer tumors: A repeatability analysis in a prospective multicentre cohort. *J. Nucl.* 2017, 58, 406–411.
 142. Orlhac, F.; Soussan, M.; Chouahnia, K.; Martinod, E.; Buvat, I. 18F-FDG PET-Derived Textural Indices Reflect Tissue-Specific Uptake Pattern in Non-Small Cell Lung Cancer. *PLoS ONE* 2015, 10, e0145063, doi:10.1371/journal.pone.0145063.
 143. Ger, R.B.; Meier, J.; Pahlka, R.B.; Gay, S.; Mumme, R.; Fuller, C.D.; Li, H.; Howell, R.M.; Layman, R.R.; Stafford, R.J.; et al. Effects of alterations in positron emission tomography imaging parameters on radiomics features. *PLoS ONE* 2019, 14, e0221877, doi:10.1371/journal.

- pone.0221877.
144. Leijenaar, R.T.H.; Carvalho, S.; Velazquez, E.R.; Van Elmpt, W.J.C.; Parmar, C.; Hoekstra, O.S.; Hoekstra, C.J.; Boellaard, R.; Dekker, A.; Gillies, R.; et al. Stability of FDG-PET Radiomics features: An integrated analysis of test-retest and inter-observer variability. *Acta Oncol.* 2013, 52, 1391–1397, doi:10.3109/0284186x.2013.812798.
 145. Prayer, F.; Hofmanninger, J.; Weber, M.; Kifjak, D.; Willenpart, A.; Pan, J.; Röhrich, S.; Langa, G.; Prosch, H. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. *Methods* 2021, 188, 98–104, doi:10.1016/j.ymeth.2020.08.007.
 146. Zhovannik, I.; Bussink, J.; Traverso, A.; Shi, Z.; Kalendralis, P.; Wee, L.; Dekker, A.; Fijten, R.; Monshouwer, R. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin. Transl. Radiat. Oncol.* 2019, 19, 33–38, doi:10.1016/j.ctro.2019.07.003.
 147. Balagurunathan, Y.; Gu, Y.; Wang, H.; Kumar, V.; Grove, O.; Hawkins, S.; Kim, J.; Goldgof, D.B.; Hall, L.O.; Gatenby, R.A.; et al. Reproducibility and Prognosis of Quantitative Features Extracted from CT Images. *Transl. Oncol.* 2014, 7, 72–87, doi:10.1593/tlo.13844.
 148. Al-Kadi, O.S. Assessment of texture measures susceptibility to noise in conventional and contrast enhanced computed to-mography lung tumour images. *Comput. Med Imaging Graph.* 2010, 34, 494–503, doi:10.1016/j.compmedimag.2009.12.011.
 149. Lee, J.; Steinmann, A.; Ding, Y.; Lee, H.; Owens, C.; Wang, J.; Yang, J.; Followill, D.; Ger, R.; MacKin, D.; et al. Radiomics feature robustness as measured using an MRI phantom. *Sci. Rep.* 2021, 11, 3973, doi:10.1038/s41598-021-83593-3.
 150. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a mul-ti-centre test-retest trial. *Sci. Rep.* 2019, 9, 48, doi:10.1038/s41598-019-41344-5.
 151. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* 2018, 102, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.
 152. Su, C.; Jiang, J.; Zhang, S.; Shi, J.; Xu, K.; Shen, N.; Zhang, J.; Li, L.; Zhao, L.; Zhang, J.; et al. Radiomics based on multicontrast MRI can precisely differentiate among glioma subtypes and predict tumour-proliferative behaviour. *Eur. Radiol.* 2019, 29, 1986–1996, doi:10.1007/s00330-018-5704-8.
 153. Han, Y.; Xie, Z.; Zang, Y.; Zhang, S.; Gu, D.; Zhou, M.; Gevaert, O.; Wei, J.; Li, C.; Chen, H.; et al. Non-invasive genotype predic-tion of chromosome 1p/19q co-deletion by development and validation of an MRI-based radiomics signature in lower-grade gliomas. *J. Neuro-Oncol.* 2018, 140, 297–306, doi:10.1007/s11060-018-2953-y.
 154. Liu, Z.; Zhang, X.-Y.; Ying-Shi, S.; Wang, L.; Zhu, H.-T.; Tang, Z.; Wang, S.; Li, X.-T.; Tian, J.; Sun, Y.-S. Radiomics Analysis for Evaluation of Pathological Complete Response to Neoadjuvant Chemoradiotherapy in Locally Advanced Rectal Cancer. *Clin. Cancer Res.* 2017, 23, 7253–7262, doi:10.1158/1078-0432.ccr-17-1038.
 155. Horvat, N.; Veeraraghavan, H.; Khan, M.; Blazic, I.; Zheng, J.; Capanu, M.; Sala, E.; Garcia-

- Aguilar, J.; Gollub, M.J.; Petkovska, I. MR Imaging of Rectal Cancer: Radiomics Analysis to Assess Treatment Response after Neoadjuvant Therapy. *Radiology* 2018, 287, 833–843, doi:10.1148/radiol.2018172300.
156. Nie, K.; Shi, L.; Chen, Q.; Hu, X.; Jabbour, S.K.; Yue, N.; Niu, T.; Sun, X. Rectal Cancer: Assessment of Neoadjuvant Chemoradiation Outcome based on Radiomics of Multiparametric MRI. *Clin. Cancer Res.* 2016, 22, 5256–5264, doi:10.1158/1078-0432.ccr-15-2997.
 157. Min, X.; Li, M.; Dong, D.; Feng, Z.; Zhang, P.; Ke, Z.; You, H.; Han, F.; Ma, H.; Tian, J.; et al. Multi-parametric MRI-based radiomics signature for discriminating between clinically significant and insignificant prostate cancer: Cross-validation of a machine learning method. *Eur. J. Radiol.* 2019, 115, 16–21, doi:10.1016/j.ejrad.2019.03.010.
 158. Viswanath, S.E.; Chirra, P.V.; Yim, M.; Rofsky, N.M.; Purysko, A.S.; Rosen, M.A.; Bloch, B.N.; Madabhushi, A. Comparing radiomic classifiers and classifier ensembles for detection of peripheral zone prostate tumors on T2-weighted MRI: A multi-site study. *BMC Med. Imaging* 2019, 19, 1–12, doi:10.1186/s12880-019-0308-6.
 159. Rai, R.; Holloway, L.C.; Brink, C.; Field, M.; Christiansen, R.L.; Sun, Y.; Barton, M.B.; Liney, G.P. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Med Phys.* 2020, 47, 3054–3063, doi:10.1002/mp.14173.
 160. Chatterjee, A.; Vallieres, M.; Dohan, A.; Levesque, I.R.; Ueno, Y.; Saif, S.; Reinhold, C.; Seuntjens, J. Creating Robust Predictive Radiomic Models for Data From Independent Institutions Using Normalization. *IEEE Trans. Radiat. Plasma Med. Sci.* 2019, 3, 210–215, doi:10.1109/trpms.2019.2893860.
 161. Haga, A.; Takahashi, W.; Aoki, S.; Nawa, K.; Yamashita, H.; Abe, O.; Nakagawa, K. Standardization of imaging features for radiomics analysis. *J. Med. Investig.* 2019, 66, 35–37, doi:10.2152/jmi.66.35.
 162. Crombé, A.; Kind, M.; Fadli, D.; Le Loarer, F.; Italiano, A.; Buy, X.; Saut, O. Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients. *Sci. Rep.* 2020, 10, 15496, doi:10.1038/s41598-020-72535-0.
 163. Nyúl, L.G.; Udupa, J.K. On standardizing the MR image intensity scale. *Magn. Reson. Med.* 1999, 42, 1072–1081, doi:10.1002/(sici)1522-2594(199912)42:63.3.co;2-d.
 164. Nyul, L.; Udupa, J.; Zhang, X. New variants of a method of MRI scale standardization. *IEEE Trans. Med. Imaging* 2000, 19, 143–150, doi:10.1109/42.836373.
 165. Wang, L.; Lai, H.-M.; Barker, G.; Miller, D.H.; Tofts, P. Correction for variations in MRI scanner sensitivity in brain studies with histogram matching. *Magn. Reson. Med.* 1998, 39, 322–327, doi:10.1002/mrm.1910390222.
 166. Masson, I.; Da-Ano, R.; Lucia, F.; Doré, M.; Castelli, J.; de Monsabert, C.G.; Ramée, J.; Sellami, S.; Visvikis, D.; Hatt, M.; et al. Statistical harmonization can improve the development of a multicenter CT-based radiomic model predictive of nonresponse to induction chemotherapy in laryngeal cancers. *Med. Phys.* 2021, 48, 4099–4109, doi:10.1002/mp.14948.
 167. Lucia, F.; Visvikis, D.; Vallières, M.; Desseroit, M.-C.; Miranda, O.; Robin, P.; Bonaffini, P.A.; Alferi, J.; Masson, I.; Mervoyer, A.; et al. External validation of a combined PET and MRI radiomics

- model for prediction of recurrence in cervical cancer patients treated with chemoradiotherapy. *Eur. J. Nucl. Med. Mol. Imaging* 2019, 46, 864–877, doi:10.1007/s00259-018-4231-9.
168. Johnson, W.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Bio-statistics* 2006, 8, 118–127, doi:10.1093/biostatistics/kxj037.
 169. Fortin, J.-P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E.; et al. Harmonization of multi-site diffusion tensor imaging data. *NeuroImage* 2017, 161, 149–170, doi:10.1016/j.neuroimage.2017.08.047.
 170. Orlhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics. *Radiology* 2019, 291, 53–59, doi:10.1148/radiol.2019182023.
 171. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J. Nucl. Med.* 2018, 59, 1321–1328, doi:10.2967/jnumed.117.199935.
 172. Da-Ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci. Rep.* 2020, 10, 10248, doi:10.1038/s41598-020-66110-w.
 173. Vetter, T.R.; Schober, P. Agreement Analysis: What He Said, She Said Versus You Said. *Anesth. Analg.* 2018, 126, 2123–2128, doi:10.1213/ane.0000000000002924.
 174. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods* 2021, 188, 20–29, doi:10.1016/j.ymeth.2020.05.022.
 175. Lin, L.I.-K. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989, 45, 255–268, doi:10.2307/2532051.
 176. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, Aaron K.; Court, L. Data from credence cartridge radiomics phantom CT scans. *Cancer Imaging Arch.* 2017, 10, K9.
 177. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS ONE* 2021, 16, e0251147, doi:10.1371/journal.pone.0251147.
 178. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.; Granzier, R.; Hustinx, R.; Mottaghy, F.; Woodruff, H.; Wildberger, J.; et al. The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization. *Cancers* 2021, 13, 1848, doi:10.3390/cancers13081848.
 179. Ibrahim, A.; Primakov, S.; Barufaldi, B.; et al. Reply to Orlhac, F.; Buvat, I. Comment on “Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* 2021, 13, 1848”. *Cancers* 2021, 13, 3080.
 180. Mackin, D.; Fave, X.; Zhang, L.; Yang, J.; Jones, A.K.; Ng, C.S.; Court, L. Harmonizing the pixel

- size in retrospective computed tomography radiomics studies. *PLoS ONE* 2017, 12, e0178524, doi:10.1371/journal.pone.0178524.
181. Rozantsev, A.; Salzmänn, M.; Fua, P. Beyond Sharing Weights for Deep Domain Adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* 2019, 41, 801–814, doi:10.1109/tpami.2018.2814042.
 182. Sun, B.; Saenko, K. Deep CORAL: Correlation Alignment for Deep Domain Adaptation. In *Computer Vision—ECCV 2016 Workshops*; Springer International Publishing: Cham, Switzerland, 2016; pp. 443–450.
 183. Sun, B.; Feng, J.; Saenko, K. Return of Frustratingly Easy Domain Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Phoenix, AZ, USA, 12–17 February 2016.
 184. Saenko, K.; Kulis, B.; Fritz, M.; Darrell, T. Adapting Visual Category Models to New Domains. In *Computer Vision—ECCV 2010*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 213–226.
 185. Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; Lempitsky, V. Domain-Adversarial Training of Neural Networks. In *Domain Adaptation in Computer Vision Applications*; Springer: Cham, Switzerland, 2017; pp. 189–209.
 186. Tzeng, E.; Hoffman, J.; Darrell, T.; Saenko, K. Simultaneous Deep Transfer across Domains and Tasks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, Istanbul, Turkey, 2–5 August 2015.
 187. Dinsdale, N.K.; Jenkinson, M.; Namburete, A.I. Deep learning-based unlearning of dataset bias for MRI harmonisation and confound removal. *NeuroImage* 2021, 228, 117689, doi:10.1016/j.neuroimage.2020.117689.
 188. Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.; O’Sullivan, B.; Aerts, H.J.; et al. Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother. Oncol.* 2019, 130, 2–9, doi:10.1016/j.radonc.2018.10.027.
 189. Rizzo, S.; Botta, F.; Raimondi, S.; Origgi, D.; Fanciullo, C.; Morganti, A.G.; Bellomi, M. Radiomics: The facts and the challenges of image analysis. *Eur. Radiol. Exp.* 2018, 2, 36, doi:10.1186/s41747-018-0068-z.

A large, white, stylized number 6 is centered on a blue, textured, watercolor-like background. The background consists of various shades of blue, from light to dark, with a mottled, organic appearance. The number 6 is a simple, clean, sans-serif font. The overall composition is abstract and artistic.

Chapter 6

The application of a workflow
integrating the variable reproducibility
and harmonizability of radiomic
features on a phantom dataset

Authors

Abdalla Ibrahim, Turkey Refaee, Ralph T.H. Leijenaar, Sergey Primakov,
Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Andrew D.A. Maidment
and Philippe Lambin

Adapted from

Plos one. 2021 May 7;16(5):e0251147

DOI

10.1371/journal.pone.0251147

Abstract

Radiomics – the high throughput extraction of quantitative features from medical images and their correlation with clinical and biological endpoints- is the subject of active and extensive research. Although the field shows promise, the generalizability of radiomic signatures is affected significantly by differences in scan acquisition and reconstruction settings. Previous studies reported on the sensitivity of radiomic features (RFs) to test-retest variability, inter-observer segmentation variability, and intra-scanner variability. A framework involving robust radiomics analysis and the application of a post-reconstruction feature harmonization method using ComBat was recently proposed to address these challenges. In this study, we investigated the reproducibility of RFs across different scanners and scanning parameters using this framework. We analysed thirteen scans of a ten-layer phantom that were acquired differently. Each layer was subdivided into sixteen regions of interest (ROIs), and the scans were compared in a pairwise manner, resulting in seventy-eight different scenarios. Ninety-one RFs were extracted from each ROI. As hypothesized, we demonstrate that the reproducibility of a given RF is not a constant but is dependent on the heterogeneity found in the data under analysis. The number (%) of reproducible RFs varied across the pairwise scenarios investigated, having a wide range between 8 (8.8%) and 78 (85.7%) RFs. Furthermore, in contrast to what has been previously reported, and as hypothesized in the robust radiomics analysis framework, our results demonstrate that ComBat cannot be applied to all RFs but rather on a percentage of those – the “ComBatable” RFs – which differed depending on the data being harmonized. . The number (%) of reproducible RFs following ComBat harmonization varied across the pairwise scenarios investigated, ranging from 14 (15.4%) to 80 (87.9%) RFs, and was found to depend on the heterogeneity in the data. We conclude that the standardization of image acquisition protocols remains the cornerstone for improving the reproducibility of RFs, and the generalizability of the signatures developed. Our proposed approach helps identify the reproducible RFs across different datasets.

Keywords

Radiomics, Harmonization, Feature stability, Feature reproducibility

Introduction

With the advancement and involvement of artificial intelligence in performing high-level tasks, its application has been extensively researched in the field of medical imaging analysis [1]. Radiomics – the high throughput extraction of quantitative features from medical imaging to find correlations with biological or clinical outcomes [2-4] – is currently one of the most commonly used quantitative imaging analysis methods in medical imaging.

A major area of research in the field of radiomics is the selection of robust and informative image features to be used as input for machine learning models [5]. Evidence suggests that radiomic features (RFs) are sensitive to differences in several factors, including make and type of imaging scanner, reconstruction settings, and protocols used to acquire the images [6, 7]. Studies on the reproducibility of RFs across test-retest [8, 9]; or across scans of a phantom made on the same scanner using different exposure levels, while fixing other parameters [10]; or across scans of a phantom using different acquisition and reconstruction parameters [11] highlighted the high sensitivity of RFs to variations within datasets.

The above-mentioned studies focused on the reproducibility of RFs in limited settings, such as test-retest, inter-observer variability, and intra-scanner variability. As these studies reported significant differences in groups of RFs, it is only intuitive that adding more variation to image acquisition and reconstruction will further dampen the reproducibility of RFs. These findings indicate that ignoring data heterogeneity will influence the performance and generalizability of the models developed, especially in studies where training and validation sets are independent. Therefore, a global initiative – the Image Biomarkers Standardization Initiative (IBSI) – has been initiated in an effort to standardize the extraction of image biomarkers (RFs) from medical images [12]. The IBSI aims to standardize both the computation of RFs and the image processing steps required before RF extraction. However, little attention has been paid in the bulk of literature to date to the heterogeneity in image acquisition and reconstruction when performing radiomics analysis. As the goal of radiomics research is to employ quantitative imaging features as clinical biomarker, the issue of accurate measurement and reproducibility must be addressed [13]. Biomarkers are defined as “the objective indications of medical state observed from outside the patient – which can be measured reproducibly”. Therefore, reproducible measurement is a corner stone in choosing a biomarker. In essence, RFs that cannot be reproduced cannot be compared or selected as biomarkers.

Combining Batches (ComBat) harmonization is a method that was introduced for

removing the effects of machinery and protocols used to extract gene expression data, in order to make gene expression data acquired at different centres comparable [14]. ComBat is a method that performs location and scale adjustments of the values presented to remove the discrepancies in RF values introduced by technical differences in the images. These sources of variation are further referred to as batch effects. ComBat was subsequently adopted in radiomics analysis, and some studies reported that ComBat outperforms other harmonization methods (e.g. histogram-matching, voxel size normalization, and singular value decomposition) in radiomics analyses [15, 16]. Several radiomics studies have reported on the successful application of ComBat in removing the differences in RFs introduced by different vendors and acquisition protocols [17-21]. These studies investigated the differences in radiomic RF distributions across different batches following the application of ComBat harmonization. In contrast to gene expression arrays, RFs have different definitions, and the batch effect might vary for each RF. Using phantom data allows one to study the variations in a given RF extracted from scans acquired with different scanners/reconstruction settings and to attribute these variations to the changes in acquisition and reconstruction, which in theory ComBat harmonization is designed to mitigate. However, we are not aware of any study that has performed a systematic evaluation of the performance of ComBat harmonization across variations between imaging parameters, which is the one of the objectives of this study.

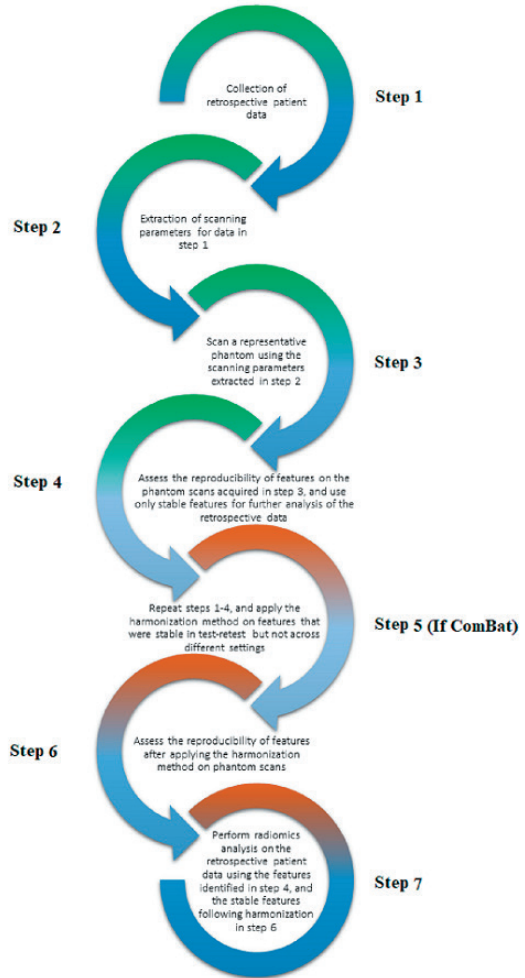


Figure 1: The proposed framework (reprinted with permission from [22]).

Ibrahim et al. (2020) have proposed a new radiomics workflow (Fig 1) that tries to address the challenges current radiomics analyses face. The framework was proposed

based on mathematical considerations of the complexity of medical imaging, and RFs' mathematical definitions. Our framework is based on the hypothesis that the reproducibility of a given RF is a not constant, but depends on the variations of image acquisition and reconstruction in the data under study. Furthermore, for ComBat to be applicable in radiomics, radiomic RF values for a given region of interest obtained after ComBat must be (nearly) identical, regardless of differences in acquisition and reconstruction.

Our general objective is to set-up the requirements for selecting biomarkers from RFs, to ease their incorporation into clinical decision support systems. We hypothesize that variations in image acquisition and reconstruction will variably affect RFs reproducibility. Furthermore, the performance of ComBat on a given RF is dependent on those variations, i.e, a given RF can be successfully harmonized with ComBat with specific variations in the imaging parameters but not others. We investigate these hypotheses on CT scans using a ten-layer radiomics phantom, which was scanned with different acquisition and reconstruction parameters on various scanner models.

Methods

Phantom Data

The publicly available Credence Cartridge Radiomics (CCR) phantom data, found in The Cancer Imaging Archive (TCIA.org) [23, 24], was used. The CCR phantom is composed of 10 different layers that correspond to further subdivided into 16 distinct

Table I: CT acquisition parameters*.

Scan	Vendor	Model	Scan Options	Effective mAs**	kVp
CCR1-001	GE	Discovery CT750 HD	HELICAL	81	120
CCR1-002	GE	Discovery CT750 HD	AXIAL	300	120
CCR1-003	GE	Discovery CT750 HD	HELICAL	122	120
CCR1-004	GE	Discovery ST	HELICAL	143	120
CCR1-005	GE	LightSpeed RT	HELICAL	1102	120
CCR1-006	GE	LightSpeed RT16	HELICAL	367	120
CCR1-007	GE	LightSpeed VCT	HELICAL	82	120
CCR1-008	Philips	Brilliance Big Bore	HELICAL	320	120
CCR1-009	Philips	Brilliance Big Bore	HELICAL	369	120
CCR1-010	Philips	Brilliance Big Bore	HELICAL	320	120
CCR1-011	Philips	Brilliance Big Bore	HELICAL	369	120
CCR1-012	Philips	Brilliance 64	HELICAL	372	120
CCR1-013	SIEMENS	Sensation Open	AXIAL	26-70	120

* Values are directly extracted from the publicly available imaging tags.

Table 2: CT acquisition parameters*.

Scan	Convolution Kernel	Filter Type	Slice thickness (mm)	Pixel spacing (mm)	kVp
CCR1-001	STANDARD	BODY FILTER	2.5	0.49	120
CCR1-002	STANDARD	BODY FILTER	2.5	0.70	120
CCR1-003	STANDARD	BODY FILTER	2.5	0.78	120
CCR1-004	STANDARD	BODY FILTER	2.5	0.98	120
CCR1-005	STANDARD	BODY FILTER	2.5	0.98	120
CCR1-006	STANDARD	BODY FILTER	2.5	0.98	120
CCR1-007	STANDARD	BODY FILTER	2.5	0.74	120
CCR1-008	B	B	3	0.98	120
CCR1-009	C	C	3	0.98	120
CCR1-010	B	B	3	1.04	120
CCR1-011	B	B	3	1.04	120
CCR1-012	B	B	3	0.98	120
CCR1-013	B31s	0	3	0.54	120

* Values are directly extracted from the publicly available imaging tags.

regions of interest (ROI) with cubic volume of 8 cm³, resulting in a total of 2080 ROIs available for further analysis. The phantom was originally scanned using 17 different imaging protocols from four medical institutes using equipment from different vendors and a variety of acquisition and reconstruction parameters. Four of the scans lacked ROI definitions, thus to maintain consistency, these were not included. The remaining 13 scans are as follows: seven different scans acquired on GE scanners, five different scans acquired on Philips scanners, and one scan acquired on a Siemens scanner (Tables 1 and 2).

Radiomic features extraction

For each ROI, quantitative imaging features were calculated using the open source Pyradiomics (V 2.0.2). The software contains IBSI-compliant RFs, with deviations highlighted in the feature definitions. For the extraction step, no changes to the original slice thickness or pixel spacing of the scans were applied. To reduce noise and computational requirements, images were pre-processed by binning voxel greyscale values into bins with a fixed width of 25 HUs prior to extracting RFs. The extracted features included HU intensity features, shape features, and texture features describing the spatial distribution of voxel intensities using 5 texture matrices (i.e., grey-level co-occurrence (GLCM), grey-level run-length (GLRLM), grey-level size-zone (GLSZM), grey-level dependence (GLDM), and neighbourhood grey-tone difference matrix (NGTDM)). Detailed description of the features can be found online at <https://pyradiomics.readthedocs.io/en/latest/features.html>.

ComBat Harmonization

ComBat employs empirical Bayes methods to estimate the differences in feature values attributed to a batch effect. Empirical Bayes methods are able to estimate the prior distribution from a given dataset via statistical inference. In the context of radiomics, ComBat assumes that feature values can be approximated by the equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \varepsilon_{ij} \quad (1)$$

where α is the average value for feature Y_{ij} for ROI j on scanner i ; X is a design matrix of the covariates of interest; β is the vector of regression coefficients corresponding to each covariate; γ_i is the additive effect of scanner i on features, which is presupposed to follow a normal distribution; δ_i is the multiplicative scanner effect, which is presupposed to follow an inverse gamma-distribution; and ε_{ij} is an error term, presupposed to be normally distributed with zero mean [17]. ComBat performs feature transformation based on the empirical Bayes prior estimates for γ and δ for each batch:

$$Y_{ij}^{ComBat} = \frac{(Y_{ij} - \hat{\alpha} - \hat{\beta}X_{ij} - \gamma_i^*)}{\delta_i^*} + \hat{\alpha} + \hat{\beta}X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β , respectively. γ_i^* and δ_i^* are the empirical Bayes estimates of γ_i and δ_i , respectively [17].

Statistical analysis

To assess the agreement of a given RF for the same ROI scanned using different settings and scanners, the concordance correlation coefficient (CCC) was calculated using epiR (version 0.9-99) [25] on R [26] (version 3.5.1), using R studio (version 1.1.456) [27]. The CCC is used to evaluate the agreement between paired readings [28], and provides the measure of concordance as a value between 1 and -1, where 0 represents no concordance, 1 represents a perfect direct positive concordance, and -1 indicates a perfect inverse concordance. It further takes into account the rank and value of the RFs. The analysis of the reproducibility before and after ComBat harmonization was performed in a pairwise manner, resulting in 78 different investigated scenarios. To assess differences in RF stability for differing data, the reproducibility of radiomics RFs across scans within a wide spectrum of scenarios was calculated. Data ranging from differences in a single acquisition or reconstruction parameter, to scans acquired using entirely different settings (See S1 table) were included. To identify reproducible radiomics, the CCC was calculated for all RFs for all ROIs across the 78 investigated scenarios. A cut-off of $CCC > 0.9$, as found in the literature, suggests that a value < 0.9 indicates poor concordance [29]. To identify the RFs that could be harmonized using ComBat, the pair-wise CCC was calculated following ComBat in each of the investigated 78

scenarios. We applied ComBat using R package “SVA” (version 3.30.1) [30]. As the RFs are calculated for the same ROI but for different scans, the agreement in RF value is expected to be high following ComBat harmonization. Thus, RFs that had a CCC<0.9 were considered to be not harmonizable with ComBat. The code used in this work is publicly available on <https://github.com/AbdallaIbrahim/The-reproducibility-and-ComBatability-of-Radiomic-features>.

Results

Reproducible Radiomic features

For each ROI, a total of 91 RFs were extracted. The number (percentage) of reproducible RFs in each pair-wise comparison ranged from 9 (8.8%) to 78 (85.7%) RFs, depending on the variations in acquisition and reconstruction of the scans (table 3). The highest concordance in feature values (85.7%) was observed between the two Philips scans (CCR1-010 and CCR1-011) that were acquired using the same scanner model, and the same acquisition and reconstruction parameters except for the effective mAs, which differed by just 15% (tables 1 and 2).

The more profound the variations in scan acquisition parameters, the smaller the concordance of the extracted RFs (tables 1-3, S1).

As stated, in the best scenario (CCR1-010 and CCR1-011), 78 (85.7%) RFs were found to be reproducible, while 13 (14.3%) RFs were found not to be reproducible. Some RFs (n=8) were found to be concordant across all pairs. These RFs were histogram-based RFs that take into account the value of a single pixel/voxel, without looking at the relationship between neighbouring pixels/voxels. These RFs are (i) original first order 10Percentile; (ii) original first order 90Percentile; (iii) original first order Maximum; (iv) original first order Mean (v) original first order Median; (vi) original first order Minimum; (vii) original first order Root Mean Squared; and (viii) original first order Total Energy. Nevertheless, the remainder (majority) of the RFs (including 10 histogram-based RFs) were not found to be reproducible across all pairs.

Looking at tables (1-3, S1), we can consider subgroups of scans. Scans CCR1-001-007 were all acquired using the same imaging vendor (GE), but different scanner models and scanning parameters. The highest number of concordant RFs in this group was found between CCR1-004 and CCR1-006 (71 RFs), which were acquired on two different scanner models, but were scanned with identical scanning parameters except for the mAs. The lowest number of concordant RFs in this group was found between scans CCR1-001 and CCR1-005 (13 RFs), which were acquired on two different scanner models, with the same scanning parameters except for the pixel spacing and mAs. Scans

Table 3: The number (percentage) of concordant RFs before ComBat harmonization between pair wise combinations of scans with different acquisition and reconstruction.

	CCRI-001	CCRI-002	CCRI-003	CCRI-004	CCRI-005	CCRI-006	CCRI-007	CCRI-008	CCRI-009	CCRI-010	CCRI-011	CCRI-012
CCRI-002	38 (41.76%)											
CCRI-003	46 (50.55%)	59 (64.84%)										
CCRI-004	18 (19.78%)	34 (37.36%)	25 (27.47%)									
CCRI-005	13 (14.29%)	23 (25.27%)	17 (18.68%)	66 (72.53%)								
CCRI-006	16 (17.58%)	24 (26.37%)	18 (19.78%)	71 (78.02%)	69 (75.82%)							
CCRI-007	49 (53.85%)	65 (71.43%)	67 (73.63%)	21 (23.08%)	14 (15.38%)	14 (15.38%)						
CCRI-008	8 (8.79%)	12 (13.19%)	14 (15.38%)	41 (45.05%)	34 (37.36%)	47 (51.65%)	10 (10.99%)					
CCRI-009	9 (9.89%)	19 (20.88%)	13 (14.29%)	67 (73.63%)	65 (71.43%)	74 (81.32%)	11 (12.09%)	48 (52.75%)				
CCRI-010	8 (8.79%)	10 (10.99%)	13 (14.29%)	32 (35.16%)	21 (23.08%)	27 (29.67%)	11 (12.09%)	59 (64.84%)	34 (37.36%)			
CCRI-011	8 (8.79%)	11 (12.09%)	12 (13.19%)	45 (49.45%)	34 (37.36%)	42 (46.15%)	11 (12.09%)	57 (62.64%)	52 (57.14%)	78 (85.71%)		
CCRI-012	8 (8.79%)	13 (14.29%)	12 (13.19%)	21 (23.08%)	16 (17.58%)	22 (24.18%)	10 (10.99%)	61 (67.03%)	36 (39.56%)	71 (78.02%)	69 (75.82%)	
CCRI-013	51 (56.04%)	44 (48.35%)	47 (51.65%)	41 (45.05%)	34 (37.36%)	32 (35.16%)	48 (52.75%)	12 (13.19%)	23 (25.27%)	10 (10.99%)	9 (9.89%)	10 (10.99%)

CCR1-007 to CCR1-012 were all acquired using one of two Philips imaging vendors. The highest number of concordant RFs is documented above. The lowest number of concordant RFs was found between CCR1-009 and CCR-010 (34 RFs), which differed in terms of the mAs, convolution kernel, filter type and pixel spacing. Looking at the group of scans that were reconstructed to the same pixel spacing (CCR1-004 to CCR1-006, CCR1-008, CCR1-009, and CCR-012), the highest number of concordant RFs was observed between CCR1-006 and CCR1-009 (74 RFs), which were acquired using two different imaging vendors, but using similar acquisition and reconstruction parameters except for the slice thickness, and kernel. The lowest number of concordant RFs was found between CCR1-005 and CCR1-012 (16 RFs), which were acquired using different imaging vendors, and different acquisition and reconstruction parameters except for the kVp. Finally, comparing scans acquired with different vendors resulted in a lower number of concordant RFs compared to scans acquired with the scanners from the same imaging vendor, except for the scenario when the majority of acquisition and reconstruction parameters were mostly identical (CCR1-006 vs CCR1-009).

ComBat harmonization

As previously shown in the literature, we used each scan as a different batch in the ComBat equation. ComBat was applied pairwise (78 different pairs) and the concordance between RFs was measured for each pair (table 4). The percentage of RFs that became concordant following ComBat application ranged from 1.4% (71 concordant RFs increased to 72) to 344% (9 concordant RFs increased to 40).

The highest number of concordant RFs following ComBat application was 80 (87.9%) RFs. In this scenario, a single acquisition parameter differed between the two scans (Philips, CCR1-010 and CCR1-011). ComBat application improved the concordance of only two RFs (80 RFs after ComBat compared to 78 RFs before), and failed to improve the concordance of the remaining 11 RFs. On the other hand, in cases where the differences in acquisition and reconstruction parameters differed more (e.g., CCR1-001 (GE) vs CCR1-007 (Philips)), the application of ComBat improved the concordance of 31 RFs, resulting in a total of 40 concordant RFs (~44% of the total number of RFs), more than 3 times the number of concordant RFs before harmonization. Furthermore, the successful application of ComBat on RFs depended on the variations in the batches defined. Only two RFs were found to be concordant in all pairwise scenarios following ComBat harmonization: (i) original first order Energy; and (ii) original gldm Small Dependence High Gray Level Emphasis; in addition to the 8 RFs mentioned above.

Table 4: The number (percentage) of concordant RFs after ComBat harmonization between pair wise combinations of scans with different acquisition and reconstruction.

	CCRI-001	CCRI-002	CCRI-003	CCRI-004	CCRI-005	CCRI-006	CCRI-007	CCRI-008	CCRI-009	CCRI-010	CCRI-011	CCRI-012
CCRI-002	63 (69.23%)											
CCRI-003	69 (75.82%)	75 (82.42%)										
CCRI-004	48 (52.75%)	72 (79.12%)	57 (62.64%)									
CCRI-005	43 (47.25%)	60 (65.93%)	54 (59.34%)	72 (79.12%)								
CCRI-006	50 (54.95%)	63 (69.23%)	59 (64.84%)	76 (83.52%)	72 (79.12%)							
CCRI-007	70 (76.92%)	69 (75.82%)	74 (81.32%)	56 (61.54%)	49 (53.85%)	57 (62.64%)						
CCRI-008	27 (29.67%)	36 (39.56%)	36 (39.56%)	61 (67.03%)	54 (59.34%)	56 (61.54%)	28 (30.77%)					
CCRI-009	40 (43.96%)	57 (62.64%)	53 (58.24%)	76 (83.52%)	74 (81.32%)	81 (89.01%)	52 (57.14%)	57 (62.64%)				
CCRI-010	18 (19.78%)	22 (24.18%)	19 (20.88%)	54 (59.34%)	48 (52.75%)	48 (52.75%)	17 (18.68%)	68 (74.73%)	53 (58.24%)			
CCRI-011	14 (15.38%)	23 (25.27%)	25 (27.47%)	67 (73.63%)	59 (64.84%)	59 (64.84%)	16 (17.58%)	65 (71.43%)	67 (73.63%)	80 (87.91%)		
CCRI-012	16 (17.58%)	29 (31.87%)	28 (30.77%)	56 (61.54%)	48 (52.75%)	49 (53.85%)	16 (17.58%)	70 (76.92%)	53 (58.24%)	72 (79.12%)	74 (81.32%)	
CCRI-013	65 (71.43%)	75 (82.42%)	69 (75.82%)	65 (71.43%)	55 (60.44%)	59 (64.84%)	67 (73.63%)	35 (38.46%)	58 (63.74%)	35 (38.46%)	36 (39.56%)	34 (37.36%)

Discussion

In this work, for our first objective to investigate RFs reproducibility, we show that the majority of RFs are affected to different amounts depending upon the variations in acquisition and reconstruction parameters. We also show that the reproducibility of a given RF is not constant, but rather it is dependent on the variations in the data under study, as seen in table 3. We identified a number of RFs that were robust to the variations in scan acquisition in the dataset we analysed. These RFs could be used without any post-processing harmonization. While the same dataset has been analysed for similar purposes previously [11, 21], we analysed the data differently, and report different results than those studies. Our results show a substantial intra-scanner variability, and even greater inter-scanner variability, which is in line with other previous findings [10, 31, 32]. Only eight RFs (~9%) of the extracted RFs showed insensitivity to the differences in acquisition shown in tables 1 and 2, and could be directly used to build radiomic signatures. The rest of the RFs (91%) could not be used without addressing the acquisition differences. Our sub-groups analysis showed that changes in pixel spacing and convolution kernel have more profound effects on the reproducibility of RFs, compared to variations limited solely to the effective mAs, scanner model or imaging vendor used. While the percentages reported are representative of the reproducibility of RFs in the data analysed, it highlights the sensitive nature of RFs, and helps set guidelines to preselect meaningful and reproducible RFs. We deduce that the use of RFs extracted from scans acquired with different hardware and parameters, without addressing the issue of reproducibility and harmonization, can lead to spurious results as the vast majority of RFs are sensitive to even minor variations in image acquisition and reconstruction. Therefore, models developed using RFs with large unexplained variances will most likely not be generalizable.

As our second aim, we investigated the applicability of ComBat harmonization to removing differences in RF values attributed to batch effects. Studies [11, 21] have reported on the reproducibility of RFs on the same or a similar dataset to the one we analysed. However, our findings and conclusions vary significantly from theirs. In contrast to previous studies, we are the first to report that the reproducibility of RFs is dependent on the variations in the data under analysis. Previous studies referred to RFs as generally reproducible or non-reproducible. Our analysis shows that a given RF can be reproducible in some scenarios and not in the others, depending on the variations in acquisition and reconstruction parameters. Moreover, ComBat was mathematically defined to remove one (technical) batch effect at a time while considering all the biologic covariates at the same time. However, as our results show (tables 3 and 4), the variations in acquisition and reconstruction parameters within one scanner, at least in some instances, have a stronger impact on the reproducibility of RFs than the variations

between two scanners. As such, grouping the scans by the scanner type is not generally the way to define “batches” in the ComBat equation [14]. In contrast to what is reported in the literature, our analysis shows ComBat did not perform uniformly on most of the RFs when there were variations in the batches being harmonized. In contrast to those studies, we employed the concordance correlation coefficient (CCC) to assess the reproducibility of RFs, since the aim of harmonization is to improve the reproducibility of data. We did not use the increment of model performance as a measure for the success of harmonization for several reasons. First, the aim of harmonization is to improve the reproducibility of RFs, and ultimately the generalizability of the developed signatures, and not their model performance [33]. Second, by assuming that an increment in the model performance following harmonization is an indication that the harmonization is successful carries with it the assumption that radiomic models decode the information under analysis; this is against the essence of the study, which is to investigate whether radiomics has that potential or not. However, by using the CCC, we ensure that the results generated are based on reproducible RFs, and are therefore generalizable, regardless of the change in model performance. Furthermore, the aim of ComBat harmonization is only to remove the variance in RF values attributed to the batch effects, while maintaining the biologic information. As such, using ComBat to correct batch effects directly on patient data without providing the correct biological covariates that actually do have an effect on RF values will lead to loss of biological signals. This is because ComBat tries to harmonize the distribution of the RF across different batches, and without providing the correct biological covariates that have effects on RF values, ComBat assumes that the variations in RF value are only attributed to the defined batch, and thus would not perform uniformly as shown in table 3. In clinical settings, this is by default spurious, as the differences in RF values are attributed to both the machine and the biology/physiology. As the aim of radiomics studies is to investigate the biological correlations of RFs, we are unable to actually provide a list of biologic covariates that influence the values. In addition, each time an observation is added to the data being harmonized, ComBat has to be re-performed, and models have to be refitted, as the estimated batch effects will change each time. Therefore, the harmonization of patient RFs should follow the process of estimating fixed batch effects on phantom data, then applying the location/scale shift estimated from the phantom data on patient data, as previously described by Ibrahim et al [22].

The pairwise approach we used shows how the variations in scan acquisition and reconstruction parameters affect the reproducibility of RFs. Therefore, aside from probably a few RFs, the reproducibility of the majority of the RFs cannot be guessed in untested scenarios. The workflow (figure 1) addresses this problem by introducing the assessment of RF reproducibility on representative phantom data. This workflow differs from existing radiomics workflows by the addition of an intermediary RF pre-

selection step between RF extraction and RF selection by one of two approaches: (i) only extracting the reproducible RFs for analysis; (ii) extracting and harmonizing the 'ComBatable' RFs before RF selection and model building. The application of ComBat and the definition of what constitutes a 'batch' should be performed based on the data being analysed, as could be deduced from tables 3 and 4. For example, RFs extracted from scans acquired with different scanner models, but similar settings were found to be more concordant than RFs extracted with the same scanner model but with profound differences in acquisition and reconstruction parameters. Our proposed radiomics analysis workflow would ensure that the RFs being analysed are not affected by scan acquisition differences, and henceforth, signatures built would be more robust and generalizable. The first part of the model (steps 1-4), where only reproducible RFs are extracted and further analysed, might significantly limit the number of RFs used for further modelling. However, using the whole framework may significantly increase the number of RFs that can be used, depending on the data under study.

While the data used for this analysis are not representative of diagnostic clinical protocols and do not provide all technical details needed for proper analysis, our aim was to show that changes in scan acquisition and reconstruction parameters differently affect the majority of RFs. The variations in the reproducibility of RFs – as well as ComBat applicability – due to the heterogeneity in acquisition and reconstruction highlight the necessity of the standardization of image acquisition and reconstruction across centres. RFs have already been reported to be sensitive to test-retest [8, 34], which is the acquisition of two separate scans using the same parameters, as well as to the variations in the parameters within the same scanner [10]. Adding the variable sensitivity of RFs to different acquisition and reconstruction parameters significantly lowers the number of RFs that could be used for the analysis of heterogeneous data. As there is currently a pressing desire to analyse big data, a sound methodology is needed to address the heterogeneity introduced by machinery in retrospective data. Nevertheless, we strongly recommend the start of imaging protocol standardization across centres to facilitate future quantitative imaging analysis.

Recently, there has been an attempt to modify ComBat methodology in radiomics analysis [35]. The authors added a modification to ComBat (B-ComBat), which adds Bootstrapping and Monte Carlo to the original ComBat. The other functionality of ComBat the authors investigated was to use one of the batches as a reference (M-ComBat). The authors compared the performance of the four versions of ComBat by comparing the performance of radiomic models developed after the use of each method. The authors reported that all the methods are equally effective [35]. Therefore, we anticipate that the modified ComBat functions will have the same limitations of the original ComBat we discussed above.

Another method to harmonize RFs that is currently gaining momentum is deep learning based harmonization. A recent study developed deep learning algorithms, which were reported to improve the reproducibility of RFs across variations in scanner type, acquisition protocols and reconstruction algorithms [36]. A more recent study [37] applied a similar approach to reduce the sensitivity of RFs to scanner types. The authors reported a significant improvement in the performance of radiomic models following harmonization. These studies highlight the potential efficacy of deep learning based harmonization methods.

One limitation of our study is in considering each scan as a separate batch effect (due to lack of data) while differences between pair batches are not similar (different numbers of varying parameters), which may have affected the performance of ComBat. Acquisition and reconstruction settings include a set of different parameters, which can singularly or collectively result in differences in RFs values. Another limitation is the lack of scans generated by other commonly used scanners and protocols in the clinics; and the lack of scans with the same settings acquired using different scanners, as the data currently available is limited to the changes introduced in the imaging parameters on the available scanners. While we did not investigate the added value of this approach on a clinical dataset, our focus in this study was in designing a framework to assess the reproducibility and ‘ComBatability’ of RFs. However, it is fair to assume that if RFs are not reproducible on phantom data, they would be equally, or possibly even more, unstable on patient datasets. For example, clinical data will be acquired at a variety of mAs values across a population of patients. Lastly, while Combat has been reported to outperform other harmonization methods in terms of apparent model performance, the systemic evaluation of the effects of these methods on the reproducibility of RFs, and the comparison with the effects of ComBat harmonization will be the aim of future studies, in addition to addressing the above mentioned limitations.

Conclusion

In conclusion, we demonstrate that the reproducibility of RFs is not a constant, but changes with variations in the data acquisition and reconstruction parameters. Moreover, ComBat cannot be successfully applied on all RFs, and its successful application on a given RF is dependent on the heterogeneity of the dataset. We conclude that ComBat harmonization should not be blindly performed on patient data, but following the estimation of adjustment parameters on a phantom dataset. We anticipate that radiomics studies will benefit from our proposed harmonization workflow, as it allows comparison of a greater number of RFs, and enhances the generalizability of radiomic models. Yet, standardization of imaging protocols remains the cornerstone for improving the generalizability of prospective quantitative image studies. We recommend the

Chapter 6

standardization of scan acquisition across centres, especially in prospective clinical trials that include medical imaging; and/or the development of a specific imaging protocols for scans acquired to be used for quantitative imaging analysis.

References

1. Walsh S, de Jong EE, van Timmeren JE, Ibrahim A, Compter I, Peerlings J, et al. Decision support systems in oncology. *JCO clinical cancer informatics*. 2019;3:1-9.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature communications*. 2014;5:4006.
3. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, Van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer*. 2012;48(4):441-6.
4. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2015;278(2):563-77.
5. Reiazi R, Abbas E, Faima P, Kwan JY, Rezaie A, Bratman SV, et al. The Impact of the Variation of Imaging Factors on the Robustness of Computed Tomography Radiomic Features: A Review. *medRxiv*. 2020.
6. van Timmeren JE, Carvalho S, Leijenaar RT, Troost EG, van Elmpt W, de Ruyscher D, et al. Challenges and caveats of a multi-center retrospective radiomics study: an example of early treatment response assessment for NSCLC patients using FDG-PET/CT radiomics. *PloS one*. 2019;14(6):e0217536.
7. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics*. 2018;102(4):1143-58.
8. van Timmeren JE, Leijenaar RT, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test–retest data for radiomics feature stability analysis: Generalizable or study-specific? *Tomography*. 2016;2(4):361.
9. Prayer F, Hofmanninger J, Weber M, Kifak D, Willenpart A, Pan J, et al. Variability of computed tomography radiomics features of fibrosing interstitial lung disease: A test-retest study. *Methods*. 2020.
10. Zhovannik I, Bussink J, Traverso A, Shi Z, Kalendralis P, Wee L, et al. Learning from scanners: Bias reduction and feature correction in radiomics. *Clinical and translational radiation oncology*. 2019;19:33-8.
11. Mackin D, Fave X, Zhang L, Fried D, Yang J, Taylor B, et al. Measuring CT scanner variability of radiomics features. *Investigative radiology*. 2015;50(11):757.
12. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv preprint arXiv:161207003*. 2016.
13. Lambin P, Leijenaar RT, Deist TM, Peerlings J, De Jong EE, Van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*. 2017;14(12):749.
14. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118-27.
15. Ligeró M, Jordi-Ollero O, Bernatowicz K, Garcia-Ruiz A, Delgado-Muñoz E, Leiva D, et al.

- Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *European radiology*. 2021;31(3):1460-70.
16. Foy JJ, Al-Hallaq HA, Grekoski V, Tran T, Guruvadoo K, Armato Iii SG, et al. Harmonization of radiomic feature variability resulting from differences in CT image acquisition and reconstruction: assessment in a cadaveric liver. *Physics in Medicine & Biology*. 2020;65(20):205008.
 17. Orlhac F, Boughdad S, Philippe C, Stalla-Bourdillon H, Nioche C, Champion L, et al. A postreconstruction harmonization method for multicenter radiomic studies in PET. *Journal of Nuclear Medicine*. 2018;59(8):1321-8.
 18. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*. 2017;161:149-70.
 19. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*. 2018;167:104-20.
 20. Orlhac F, Humbert O, Boughdad S, Lasserre M, Soussan M, Nioche C, et al. Validation of a harmonization method to correct for SUV and radiomic features variability in multi-center studies. *Journal of Nuclear Medicine*. 2018;59(supplement 1):288-.
 21. Orlhac F, Frouin F, Nioche C, Ayache N, Buvat I. Validation of a method to compensate multicenter effects affecting CT Radiomics. *Radiology*. 2019:182023.
 22. Ibrahim A, Primakov S, Beuque M, Woodruff H, Halilaj I, Wu G, et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. *Methods*. 2020.
 23. Mackin DF, Xenia; Zhang, Lifei; Fried, David; Yang, Jinzhong; Taylor, Brian; Rodriguez-Rivera, Edgardo; Dodge, Cristina; Jones, Aaron Kyle; and Court, Laurence. Data From Credence Cartridge Radiomics Phantom CT Scans. The Cancer Imaging Archive. 2017. doi: <http://doi.org/10.7937/K9/TCIA.2017.zuzrml5b>.
 24. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Journal of digital imaging*. 2013;26(6):1045-57.
 25. Stevenson M, Nunes T, Sanchez J, Thornton R, Reiczigel J, Robison-Cox J, et al. *epiR: An R package for the analysis of epidemiological data*. R package version 09-43. 2013.
 26. Team RC. *R: A language and environment for statistical computing*. 2013.
 27. Team R. *RStudio: Integrated Development for R*. Boston: RStudio Inc.; 2015. 2016.
 28. Lawrence I, Lin K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;255-68.
 29. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA Client Report: HAM2005-062. 2005.
 30. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The *sva* package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882-3.
 31. Kim H, Park CM, Lee M, Park SJ, Song YS, Lee JH, et al. Impact of reconstruction algorithms on CT radiomic features of pulmonary tumors: analysis of intra-and inter-reader variability and inter-reconstruction algorithm variability. *PLoS One*. 2016;11(10):e0164924.

32. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing agreement between radiomic features computed for multiple CT imaging settings. *PloS one*. 2016;11(12):e0166550.
33. Vetter TR, Schober P. Agreement analysis: what he said, she said versus you said. *Anesthesia & Analgesia*. 2018;126(6):2123-8.
34. Leijenaar RT, Carvalho S, Velazquez ER, Van Elmpt WJ, Parmar C, Hoekstra OS, et al. Stability of FDG-PET radiomics features: an integrated analysis of test-retest and inter-observer variability. *Acta oncologica*. 2013;52(7):1391-7.
35. Da-ano R, Masson I, Lucia F, Doré M, Robin P, Alfieri J, et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Scientific Reports*. 2020;10(1):10248. doi: 10.1038/s41598-020-66110-w.
36. Andrearczyk V, Depaepe A, Müller H. Neural network training for cross-protocol radiomic feature standardization in computed tomography. *Journal of Medical Imaging*. 2019;6(2):024008.
37. Marcadent S, Hofmeister J, Preti MG, Martin SP, Van De Ville D, Montet X. Generative Adversarial Networks Improve the Reproducibility and Discriminative Power of Radiomic Features. *Radiology: Artificial Intelligence*. 2020;2(3):e190035.



7a

Chapter 7a

The effects of in-plane spatial resolution on CT-based radiomic features' stability with and without ComBat harmonization

Authors

Abdalla Ibrahim, Turkey Refaee, Ralph T.H. Leijenaar, Sergey Primakov, Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Andrew D.A. Maidment and Philippe Lambin

Adapted from

Cancers. 2021 Jan;13(8):1848

DOI

10.3390/cancers13081848

Abstract

While handcrafted radiomic features (HRFs) have shown promise in the field of personalized medicine, many hurdles hinder its incorporation into clinical practice, including but not limited to their sensitivity to differences in acquisition and reconstruction parameters. In this study, we evaluated the effects of differences in in-plane spatial resolution (IPR) on HRFs, using a phantom dataset (n=14) acquired on two scanner models. Further, we assessed the effects of interpolation methods (IMs), the choice of a new unified in-plane resolution (NUIR), and ComBat harmonization on the reproducibility of HRFs. The reproducibility of HRFs was significantly affected by variations in IPR, with pairwise concordant HRFs, as measured by the concordance correlation coefficient (CCC), ranging from 42% to 95%. The number of concordant HRFs (CCC > 0.9) after resampling varied depending on (i) the scanner model, (ii) the IM, and (iii) the NUIR. The number of concordant HRFs after ComBat harmonization depended on the variations between the batches harmonized. The majority of IMs resulted in a higher number of concordant HRFs compared to ComBat harmonization, and the combination of IMs and ComBat harmonization did not yield a significant benefit. Our developed framework can be used to assess reproducibility and harmonizability of RFs.

Keywords

Image Processing, Harmonization, Reproducibility, Radiomics biomarkers

Introduction

In recent years, quantitative medical imaging research using handcrafted radiomic features (HRFs) has been growing exponentially [1,2]. Radiomics refers to the high throughput extraction of quantitative imaging features that are expected to correlate with clinical and biological characteristics of patients [3,4]. For decades, it has been hypothesized that image texture analysis could potentially extract more information from an ROI than that solely perceived by the human eye [5,6]. Yet, the term radiomics has only been introduced recently [7,8]. HRFs are generally grouped into shape, intensity, and textural features. To date, many studies have reported on the potential of radiomics to predict various clinical endpoints [9,10]. However, major challenges, including the reproducibility of the HRFs across different acquisition and reconstruction parameters, have hindered the incorporation of radiomics in clinical decision support systems [11,12].

The essence of radiomics is that certain HRFs help decode biologic information [8], allowing these features to be treated as biomarkers. The mainstay of a biomarker is the ability to quantify it in a reproducible manner [13]. HRFs are mathematical equations applied to numeric arrays of intensity values which form the medical image. Therefore, it is intuitive that changes in the values in the array (due to differences in scan acquisition and reconstruction parameters), by the transitive property, lead to (potentially significant) quantitative changes in the HRFs. It is well established that changes in scan acquisition and reconstruction parameters affect the values in the array representing the medical image [14]. Therefore, it is a common clinical practice to scan a phantom to calibrate the CT scanner on a routine basis. Hence, similar practices are needed before radiomics studies are conducted, when the scans under analysis were acquired using heterogeneous acquisition and reconstruction parameters [15]. Many studies have already reported on the sensitivity of HRFs to different factors including: (i) temporal variability, or test-retest [16,17], in which two scans of a patient (or a phantom) are taken after a time interval using the exact scanning parameters; (ii) scanning parameters variability [11,18,19], in which an object (usually a phantom) is scanned multiple times using different scanning parameters. Variations in the majority of scanner/scanning parameter combinations were reported to impact the reproducibility of HRFs significantly [18-20].

One scan reconstruction parameter expected to have an effect on the reproducibility of HRFs is the in-plane spatial resolution (IPR), which is dictated in part by the pixel dimensions, while the through-plane spatial resolution is determined by the slice thickness and slice spacing. Resampling all the scans in a data set to a new unified in-plane spatial resolution (NUIR) before feature extraction has been employed as a method to reduce the variation in radiomic feature values [21,22]. The NUIR is usually

decided based on the most frequent IPR in the dataset and different interpolation methods (IMs) can be used for this purpose. Interpolation is a model-based method to recover continuous data from discrete data within a known range of data spacings (i.e., pixel size in images) [23]. The degree to which data recovery is possible is highly sensitive to the interpolation method and the underlying data structure. In the case of medical imaging analysis, interpolation is employed either to convert the spatial sampling rate (measured in pixel or voxel count per unit of length per dimension) to another, or to distort the image in the case of image registration [24]. Since the vast majority of HRFs are derived from pixel/voxel values and their distributions, interpolation to a common pixel spacing could potentially reduce variance introduced to these HRFs arising from differences in IPR.

As a rule, one must distinguish between interpolation methods that increase or reduce the image resolution. Interpolation from smaller pixels to larger pixels (i.e. reducing spatial resolution) usually involves some form of averaging, with the possible exception of modern deep learning-based methods.

Generally, while data acquired with small pixels will contain more noise, the process of averaging to large pixels will ameliorate the noise properties. As such, the process is less sensitive to the interpolation method/model. Interpolation from larger pixels to smaller pixels (i.e. increasing spatial resolution) on the other hand is fraught with challenges as the interpolated data can be highly sensitive to the interpolation model due to the need to create *de novo* pixel values. Larger pixels average the signal over a larger area than smaller ones, leading to the loss of variations in the original scene that occur over spatial frequencies smaller than the Nyquist limit and cannot be recovered exactly.

Certain methods, such as nearest neighbour interpolation (also called pixel replication), while fast, are less accurate than other methods such as sinc interpolation or deep-learning methods (which are trained with representative data). However, all such interpolation methods are sensitive to biases arising from the image [25]. The application of these methods to medical imaging has been evaluated qualitatively [26]. Yet, the effects of these methods on the reproducibility of HRFs is not well understood. Unlike humans, whose exposure to a vast assortment of scanners, patients, and acquisition conditions (including IPR) leads to a tolerance for such changes, IPR is likely to have more profound effects on HRFs.

A harmonization method that has become increasingly common in the field of radiomics is ComBat. ComBat was originally developed for the harmonization of gene expression arrays [27]. Several studies have investigated the potential of ComBat in radiomics analysis and recommended its use [28,29]. We hypothesize that ComBat, the

chosen IM, and the selected NUIR will affect the reproducibility of HRFs differently. In this study, the reproducibility of HRFs was assessed across different IPRs, while keeping all other parameters fixed, using a public dataset of CT scans of a phantom. A thorough investigation of the applicability of 10 different IMs was performed in an effort to identify suitable IMs for the purpose of increasing the number of reproducible HRFs in a heterogeneous dataset. In particular, we investigated whether data with discordant pixel sizes need to be interpolated to a common pixel size to perform radiomics analysis, and how the choice of IM and NUIR, as well as ComBat harmonization, affect the reproducibility of HRFs. Furthermore, we developed a generalizable workflow that assesses the impact of different harmonization techniques (Figure 1) on the reproducibility of RFs. Ultimately, the goal of our work is to guide robust radiomics analysis to ease its incorporation in clinical decision-making.

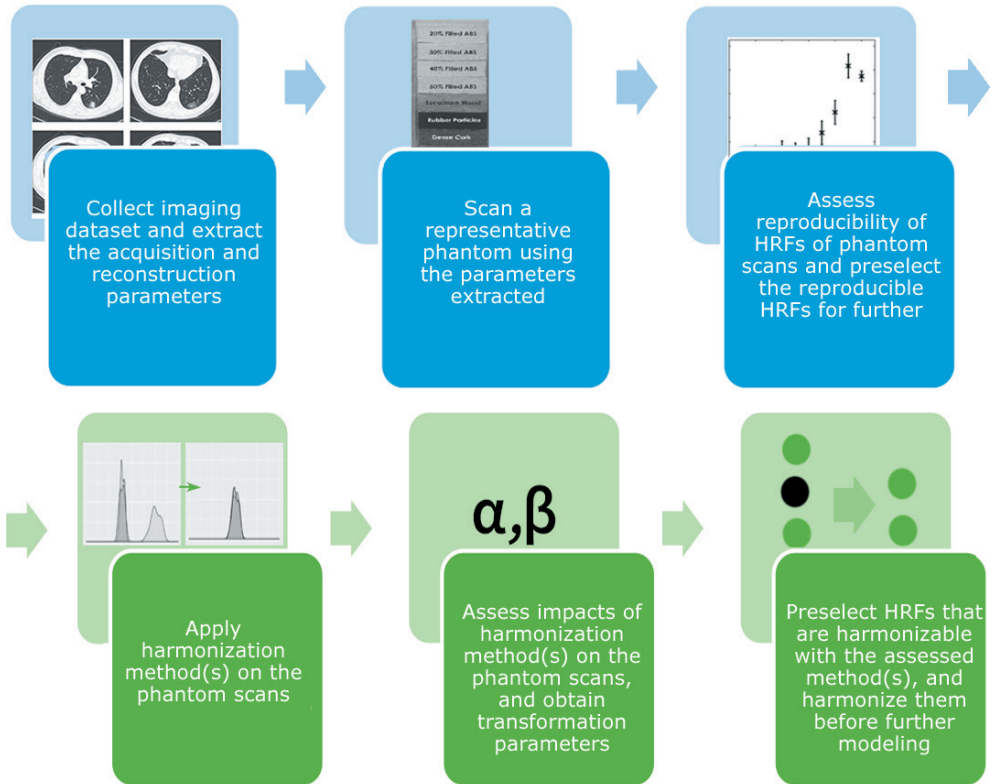


Figure 1: Proposed reproducible radiomic analysis workflow.

Materials and Methods

Phantom data

The publicly available Credence Cartridge Radiomics (CCR) phantom data [30] found in The Cancer Imaging Archive (TCIA.org) [31] was used. The CCR phantom is composed of 10 different layers that correspond to different texture patterns spanning a range of almost -900 to +700 HU (Figure S1). The publicly available dataset includes 251 scans of the phantom acquired using six scanner models manufactured by three different manufacturers. The scans were acquired using various acquisition and reconstruction parameters to assess the reproducibility of HRFs. For the purpose of this study, 14 scans acquired using 2 different scanner models (Discovery STE & LightSpeed Pro 32) of the same manufacturer (GE), which were all acquired at a single slice thickness (1.25 mm), tube voltage (120 kV), tube current (250 mA), and convolution kernel (standard), but varying IPR (Table 1) were used. The reasoning behind this selection is multifold: (i) the effects of the variations are expected to be dependent on the heterogeneity in acquisition; (ii) the number and complexity of the different combinations available are too huge to be described, analyzed and presented in a single experiment; (iii) the data under analysis were acquired using the same scanner models, and the same acquisition and reconstruction parameters except for the in-plane resolution, which allows the assessment of the effect of variations in this single parameter.

Table I: Scanning parameters of the phantom data.

Scanner		Pixel spacing (mm ²)
Discovery STE	LightSpeed Pro 32	
CCR-2-001	CCR-2-022	0.39*0.39
CCR-2-002	CCR-2-023	0.49*0.49
CCR-2-003	CCR-2-024	0.59*0.59
CCR-2-004	CCR-2-025	0.68*0.68
CCR-2-005	CCR-2-026	0.78*0.78
CCR-2-006	CCR-2-027	0.88*0.88
CCR-2-007	CCR-2-028	0.98*0.98

Interpolation and image resampling

The effects of the IMs included in the popular open-source radiomics toolbox PyRadiomics [33] were assessed in this study. The methods are based on the python library Simple-ITK [33], and include (i) nearest neighbour (NN), (ii) linear, (iii) basis spline (B-spline), (iv) Gaussian, (v) Gaussian using labelling (mask) information (LabelGaussian), and windowed sinc interpolations using the following window types: (vi) Hamming (HammingWindowedSinc or HWS), (vii) Cosine (CosineWindowedSinc or CWS), (viii) Welch (WelchWindowedSinc or WWS), (ix) Lanczos window (LanczosWindowedSinc or LWS), and (x) Blackman (BlackmanWindowedSinc or BWS).

The simplest of these IMs, and the ones with the lowest computational costs, are (i) the NN interpolation, which functions by assigning any new voxel the same value as its closest neighbor in the original image; and (ii) linear interpolation, in which the values of new pixels are interpolated linearly between the two original values [26]. B-spline interpolation is more complex than NN or linear; the calculations span four pixels [34]. While the method performs well in terms of radiologic evaluation in which the aim is to convince human observers, it is known to unnecessarily over-smooth the image [26]. The windowed sinc functions are complex convolution based interpolations that are based on multiplying the sinc function by a limited spatial support window to reduce unwanted effects on the resampled image [35], followed by filtering of the frequencies to avoid the injection of spurious frequency components. Windowed sinc functions are generally considered superior to other interpolation methods as little superfluous noise is injected into the interpolated images.

HRFs extraction

Each scan contained 10 independent regions of interest (ROIs) (one for each layer of the phantom) that occupy the same physical area of the phantom on each scan. For each ROI, HRFs were calculated using the open source software Pyradiomics V 2.1.2. HRFs were extracted multiple times to perform different experiments. First, to assess the effect of differences in in-plane resolution and ComBat harmonization on HRFs, no changes to the original in-plane resolution were made. Second, to assess the effect of different IMs and NUIRs and the combination of interpolation and ComBat, HRFs were extracted from the scans using all IMs and all available NUIRs in the dataset (Table 1).

For each set of scans (7 scans, with 10 ROIs per scan) from each scanner model (n=2), HRFs were extracted 71 times. The HRFs were extracted one time from the original scans, and 70 times with unique combinations of IM and NUIR. In each run, a total of 91 original RFs were extracted. In Pyradiomics, shape features are calculated on the original input image, and are not affected by the in-application resampling. Therefore,

those HRFs were excluded.

To reduce noise and computational requirements, images were pre-processed by binning voxel grayscale values into bins with a fixed width of 25 HUs for extracting HRFs from unfiltered images. No other image pre-processing steps were performed. The extracted HRFs included HU intensity features, and texture features describing the spatial distribution of voxel intensities using 5 texture matrices (grey-level co-occurrence (GLCM), grey-level run-length (GLRLM), grey-level size-zone (GLSZM), grey-level dependence (GLDM), and neighborhood grey-tone difference (NGTDM) matrices). A more detailed description of the Pyradiomics HRFs can be found online (<https://pyradiomics.readthedocs.io/en/latest/features.html>).

ComBat harmonization

ComBat is an empirical Bayes based method used to estimate the effects of different batches on HRFs; in this scenario, variations in scan acquisition and reconstruction parameters were considered [27]. ComBat method assumes that a feature value can be approximated by the equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \epsilon_{ij} \quad (1)$$

where α is the average value for feature Y_{ij} for ROI j on scanner i ; X is a design matrix of the biologic covariates known to affect the HRFs; β is the vector of regression coefficients corresponding to each biologic covariate; γ_i is the additive effect of scanner i on HRFs, δ_i is the multiplicative scanner effect, and ϵ_{ij} is an error term, presupposed to be normally distributed with zero mean. Based on the values estimated, ComBat performs feature transformation in the form of:

$$Y_{ij}^{ComBat} = \frac{(Y_{ij} - \hat{\alpha} - \hat{\beta} X_{ij} - \gamma_i^*)}{\delta_i^*} + \hat{\alpha} + \hat{\beta} X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of parameters α and β , respectively. γ_i^* and δ_i^* are the empirical Bayes estimates of γ_i and δ_i , respectively [28].

Statistical analysis

To assess the agreement of a given HRF for the same ROI scanned using different settings and scanners, the concordance correlation coefficient (CCC) was calculated using the epiR package (Version 0.9-99) [36] and R language (Version 3.5.1) [37] with R studio (Version 1.1.456) [38]. The CCC is used to evaluate the agreement between paired readings [38], and provides the measure of concordance as a value between 1 and -1, where 0 represents no concordance and 1 or -1 represent a perfect direct positive or inverse concordance, respectively. The CCC metric further has the advantages of (i)

robustness in small sample sizes, and (ii) taking the rank and value of the feature into consideration [39]. The cut-off of ($CCC > 0.9$) was used to select reproducible HRFs, as the literature suggests that values < 0.9 indicate poor concordance [40].

Four different approaches for assessing concordances of HRFs were used (Figure 2): (i) HRFs extracted from the original scans; (ii) HRFs extracted from the original scans and harmonized using ComBat; (iii) HRFs extracted from resampled scans; and (iv) HRFs extracted from resampled scans harmonized using ComBat. For (i), the CCC was calculated for all HRFs of all ROIs across 7 different scans from each scanner. In each run, the CCC was calculated between a different pair of scans. For (ii), HRFs with nearly zero variance (i.e HRFs which have the same value in 95% or more of the data points) had to be removed before applying ComBat. Parametric prior estimations were used, and no reference batch was assigned for ComBat application. The CCC was calculated after harmonizing the remaining HRFs using ComBat. In each run, ComBat was applied on two batches (scans). For (iii), the CCC was calculated for the HRFs following feature extraction with each of the IMs. The effects of the NUIR were assessed by calculating the CCC for the HRFs after resampling all the scans to one of the available in-plane resolutions. For (iv), ComBat was applied after the same process in (iii), and the CCC was then calculated. To gauge an overall image of the reproducibility of HRFs across all pairs as well as the impact of IMs, NUIRs, and ComBat, the number (percentage) of HRFs that were reproducible by taking the intersection of HRFs that were reproducible in each pairwise comparison of a certain scenario were compared (21 pairs in each scenario as shown in tables 2-5).

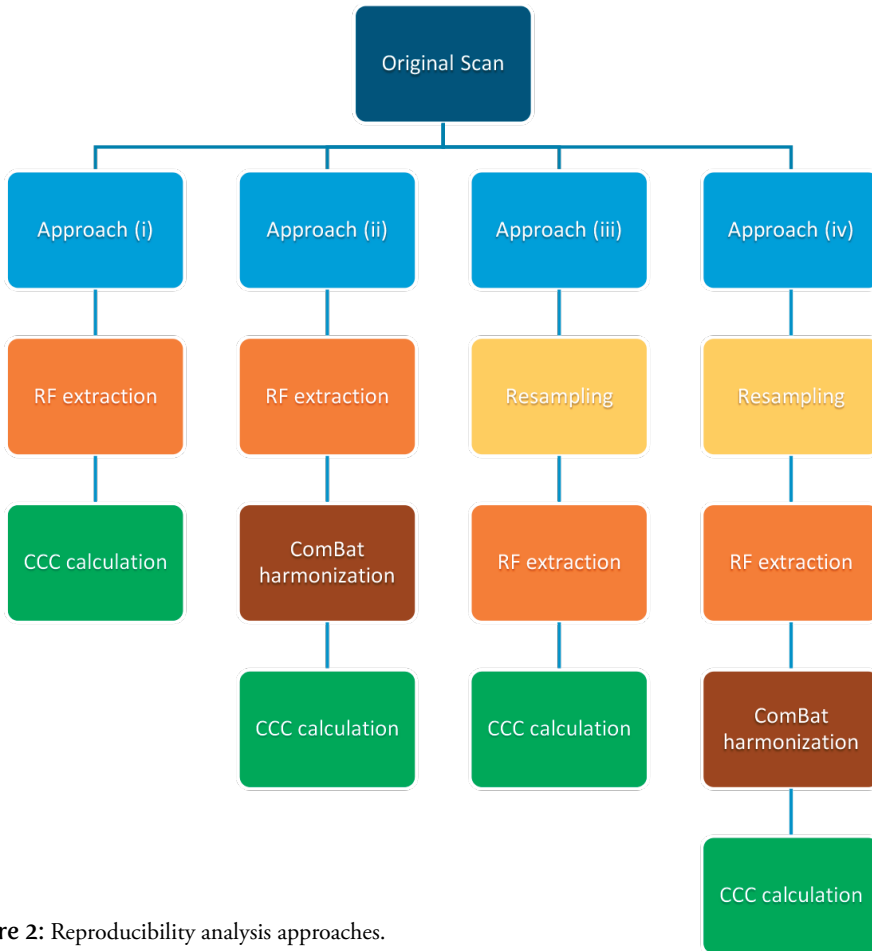


Figure 2: Reproducibility analysis approaches.

Further, we assessed the correlation between the HRFs that were concordant across all pairwise comparisons on each scanner model, using Spearman correlation [42]. HRFs were considered highly correlated if the Spearman’s correlation coefficient had a value > 0.90.

Results

Approach (i): Effects of IPR on the reproducibility of HRFs

The number of HRFs insensitive to the variations in IPR depended on the scanner model (Tables 2 and S1). In pairwise comparisons, the number of concordant HRFs was lower when the difference in IPR between the scan pairs was greater. The lowest con-cordance was observed between the scan with the highest resolution and the scan with the lowest resolution.

Out of the 91 extracted HRFs, between 39 (42.9%) and 86 (94.5%) HRFs were concordant, varying pairwise and scanner wise. Some HRFs were robust to variations in IPR in one scanner model, and not in the other.

Table 2: Number of pair-wise concordant HRFs with a CCC > 0.9 before resampling, Discovery STE model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	75 (82.4%)					
CCR-2-003	57 (62.6%)	78 (85.7%)				
CCR-2-004	53 (58.2%)	64 (70.3%)	83 (91.2%)			
CCR-2-005	50 (54.9%)	61 (67.0%)	72 (79.1%)	86 (94.5%)		
CCR-2-006	51 (56.0%)	58 (63.7%)	68 (74.7%)	76 (83.5%)	85 (93.4%)	
CCR-2-007	39 (42.9%)	42 (46.2%)	44 (48.4%)	52 (57.1%)	60 (64.9%)	83 (91.2%)

On the Discovery STE model (GE), the number of concordant HRFs ranged between 39 (42.9%) and 86 (94.5%), with a median of 70 (39.6%) HRFs (Table 2). 36 (39.6%) HRFs were reproducible regardless of the IPR selected when all other scanning parameters were fixed (List S1). Of these 36 HRFs, nine remained after removing highly correlated HRFs (List S3), and none was highly correlated with volume. Overall, the Lightspeed Pro 32 model showed lower concordance than the Discovery STE model. The number of pairwise concordant HRFs on the Lightspeed Pro 32 model ranged between 39 (42.8%) and 82 (90.1%), with a median of 60 (65.9%) (Table S1). 27 (29.7%) HRFs were reproducible across all pairs (List S2). Of these 27 HRFs, nine remained after removing highly correlated HRFs (List S4), and none was highly correlated with volume. 26 (28.6%) HRFs were reproducible on both scanner models regardless of the IPR.

Approach (ii): ComBat harmonization of HRFs extracted from original scans

ComBat harmonization increased the number of concordant HRFs compared to before harmonization. On the Discovery model, the increment in the number (percentage) of HRFs ranged between 0 (0%) and 13 (14.3%), with a median of 6 (6.6%) of the total depending on the batches being harmonized (Table 3). 46 (50.5%) HRFs were found to be reproducible across all pairwise comparisons following ComBat harmonization, 35 of which were found to be highly correlated. The number of concordant HRFs decreased with the increment in IPR variation. Hence, the increment in the number of concordant HRFs was larger when the batches being harmonized had a larger difference in IPR.



Table 3: Number of pair-wise concordant HRFs with a CCC > 0.9 after ComBat harmonization, Discovery STE model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	79 (86.8%)					
CCR-2-003	65 (71.4%)	79 (86.8%)				
CCR-2-004	59 (64.8%)	70 (76.9%)	83 (91.2%)			
CCR-2-005	58 (63.7%)	66 (72.5%)	75 (82.4%)	87 (95.6%)		
CCR-2-006	57 (62.6%)	65 (71.4%)	70 (76.9%)	84 (92.3%)	86 (94.5%)	
CCR-2-007	48 (52.7%)	55 (60.4%)	57 (62.6%)	60 (65.9%)	73 (80.2%)	84 (92.3%)

The performance of ComBat had a similar pattern on both the Discovery STE and the Lightspeed Pro 32 models. The increment in the number (percentage) of concordant HRFs extracted from the scans acquired with the Lightspeed Pro 32 model following ComBat harmonization ranged between 1 (1.1%) and 14 (15.4%) HRFs with a median increment of 7 (7.7%) HRFs compared to before harmonization, depending on the batches being harmonized (Table S2). 41 (45.1%) HRFs were reproducible across all pairs following ComBat harmonization, 29 of which were found to be highly correlated.

Approach (iii): The effects of different IMs and NUIR on HRFs

Different interpolation methods showed different effects on the reproducibility of HRFs. These effects further depended on the selected NUIR and the scanner model (Figures 3 and S2). For the majority of combinations of scanner models, IMs and NUIRs, some HRFs were only concordant when extracted from the original scans, some HRFs became concordant only after resampling, while some lost their concordance following resampling (tables S5 and S6). CSW resampling to the highest and lowest resolutions are used below as detailed examples on both scanner models.

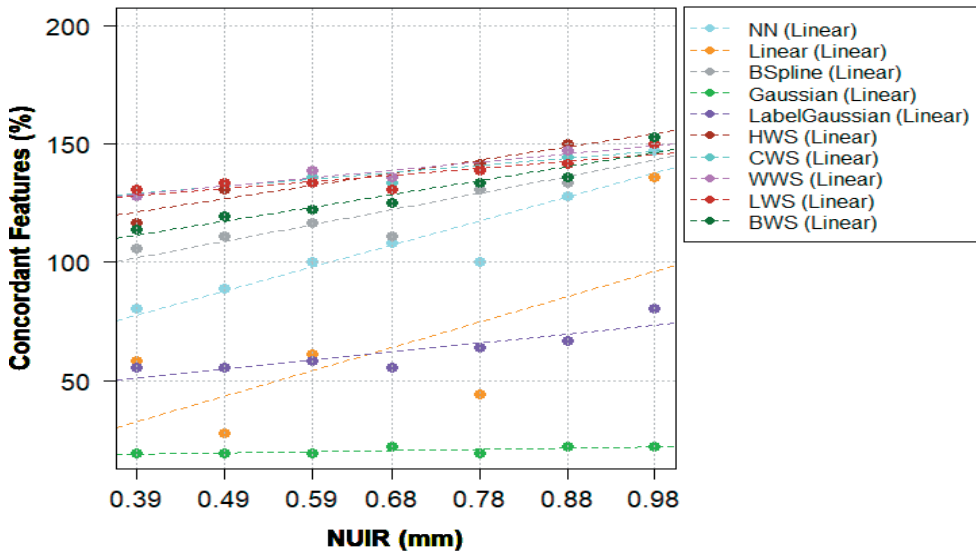


Figure 3: The percentage of concordant HRFs following resampling compared to no resampling with linear trendlines, Discovery STE model.

The performance of ComBat had a similar pattern on both the Discovery STE and the Lightspeed Pro 32 models. The increment in the number (percentage) of concordant HRFs extracted from the scans acquired with the Lightspeed Pro 32 model following ComBat harmonization ranged between 1 (1.1%) and 14 (15.4%) HRFs with a median increment of 7 (7.7%) HRFs compared to before harmonization, depending on the batches being harmonized (Table S2). 41 (45.1%) HRFs were reproducible across all pairs following ComBat harmonization, 29 of which were found to be highly correlated.

Approach (iii): The effects of different IMs and NUIR on HRFs

Different interpolation methods showed different effects on the reproducibility of HRFs. These effects further depended on the selected NUIR and the scanner model (Figures 3 and S2). For the majority of combinations of scanner models, IMs and NUIRs, some HRFs were only concordant when extracted from the original scans, some HRFs became concordant only after resampling, while some lost their concordance following resampling (tables S5 and S6). CSW resampling to the highest and lowest resolutions are used below as detailed examples on both scanner models.

7a

Table 4: Number of pair-wise concordant HRFs with a CCC > 0.9 after resampling* using CWS, Discovery model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-002	89 (97.8%)					
CCR-2-003	86 (94.5%)	88 (96.7%)				
CCR-2-004	86 (94.5%)	85 (93.4%)	88 (96.7%)			
CCR-2-005	86 (94.5%)	88 (96.7%)	91 (100%)	89 (97.8%)		
CCR-2-006	78 (85.7%)	77 (84.6%)	83 (91.2%)	79 (86.8%)	88 (96.7%)	
CCR-2-007	53 (58.2%)	53 (58.2%)	55 (60.4%)	54 (59.3%)	60 (65.9%)	85 (93.4%)

* All scans were resampled to the median pixel spacing value (0.49*0.49 mm²).

HWS performed the best when the images were resampled to a NUIR equal to or lower than the median (0.49*0.49 mm²), while CWS, WWS and LWS methods performed better on NUIR values higher than the median. BSpline IM resulted in a minor to significant increment in the number of reproducible HRFs, with higher number of concordant features when higher NUIRs were chosen. Gaussian and Label-Gaussian IMs consistently resulted in lower numbers of concordant HRFs. The number of HRFs losing concordance across all pairs when using a Gaussian IM ranged between -29 (-31.9%) and -30 (-33%) HRFs, while the range for LabelGaussian was between -11 (-12.1%) and -19 (-20.9%) HRFs, depending on the NUIR. The rest of IMs (NN and Linear) resulted in an overall decrease in the number of concordant HRFs when a NUIR below the median resolution was selected, and a minor-significant improvement with NUIRs higher than the median resolution (Table S5).

On the Lightspeed Pro 32 model, windowed sinc IMs (except for BWS) showed a consistent increment in the number of reproducible HRFs, and varying depending on the NUIR. When scans were resampled to the highest resolution using CWS, the increment in the number of concordant HRFs ranged between -9 (-9.9%) and 36 (39.6%), with a median of 8 (8.8%) HRFs. 30 (33%) HRFs were concordant across all pairs. When scans were resampled to the lowest resolution using CWS, the increment in the number of concordant HRFs ranged between -3 (-3.3%) and 31 (34.1%), with a median of 16 (17.6%) HRFs. 38 (41.8 %) HRFs were concordant across all pairs. Table S3 shows the pairwise number (percentage) of concordant HRFs following resampling to the median IPR value with CWS IM on the LightSpeed Pro 32 model, for comparison with table S4. The application of other IMs (BWS, NN, Linear, Gaussian and Label-Gaussian) with a NUIR other than the two lowest resolutions available resulted in an overall decrease in the number of concordant HRFs. However, when the lowest resolution was selected as NUIR, BSpline IM outperformed all other methods when the number of concordant HRFs across all pairs was considered (Table S6).

Approach (iv): The combination of IMs and ComBat harmonization

Approach (iii) resulted in a higher number of concordant HRFs in the majority of pairwise scenarios compared to approach (ii) for the majority of IMs that performed solely well (for example, table 3 vs table 4). The application of ComBat harmonization on HRFs extracted from resampled scans varied per scanner model, IMs, NUIRs, and batches. However, when the number of concordant HRFs across all pairs is considered, ComBat increased the number of concordant HRFs in almost all of the investigated scenarios (Figures 4 and S3; tables S7 and S8).

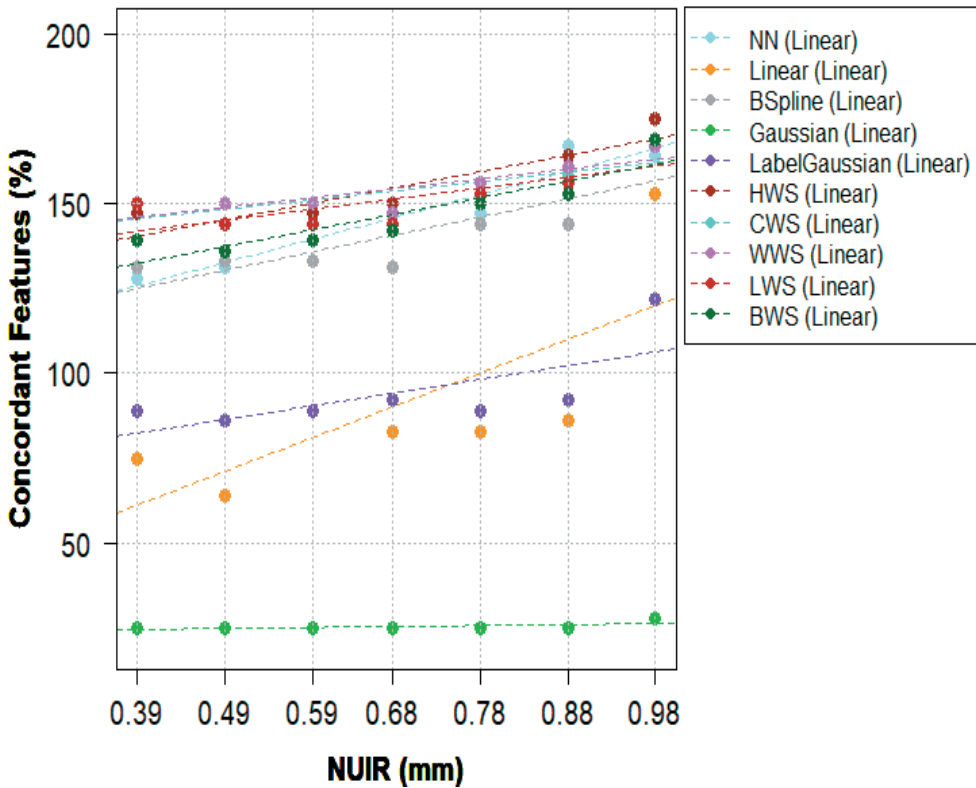


Figure 4: The percentage of concordant HRFs following resampling and ComBat harmonization compared to no resampling with linear trendlines, Discovery STE model.

On the Discovery model, the increment in the number (percentage) of concordant HRFs extracted from scans resampled to the highest resolution after ComBat harmonization ranged between 0 (0%) and 10 (11%), with a median increment of 0 (0%) of the total number of HRFs compared to before harmonization. 54 (59.3%) HRFs were concordant across all pairs. When ComBat was applied on HRFs extracted from scans resampled to the lowest resolution, the increment in the number (percentage) of HRFs ranged between -1 (-1.1%) and 10 (11%) HRFs, with a median of 0 (0%), depending

7a

on the batches being harmonized. 61 (67%) were found to be stable across all pairs. Table 5 shows the Number of pair-wise concordant HRFs following the application of ComBat on scans acquired on the Discovery STE model, and resampled to the median IPR value using CWS IM.

Table 5: Number of pair-wise concordant HRFs with a CCC > 0.9 after ComBat following resampling* us-ing CWS, Discovery STE model.

Scan	CCR-2-001	CCR-2-002	CCR-2-003	CCR-2-004	CCR-2-005	CCR-2-006
CCR-2-023	89 (97.8%)					
CCR-2-024	86 (94.5%)	88 (96.7%)				
CCR-2-025	86 (94.5%)	85 (93.4%)	88 (96.7%)			
CCR-2-026	86 (94.5%)	88 (96.7%)	91 (100%)	89 (97.8%)		
CCR-2-027	79 (86.8%)	78 (85.7%)	84 (92.3%)	84 (92.3%)	89 (97.8%)	
CCR-2-028	57 (62.6%)	61 (67.0%)	60 (65.9%)	59 (64.8%)	72 (79.1%)	85 (93.4%)

* All scans were resampled to the median pixel spacing value (0.49*0.49 mm²).

On the LightSpeed Pro 32 model, the increment in the number (percentage) of concordant HRFs after ComBat harmonization on HRFs extracted from scans resampled to the highest resolution (lowest concordance) ranged between -1 (-1.1%) and 13 (14.3%) HRFs, with a median of 3 (3.3%) of the total number of HRFs compared to before harmonization. 42 (46.2%) HRFs were concordant across all pairs. When ComBat was applied on HRFs extracted from scans resampled to the lowest resolution (highest concordance), the increment in the number (percentage) of HRFs ranged between 0 (0%) and 10 (11%) HRFs, with a median increment of 1 (1.1%) feature. 51 (56%) HRFs were concordant across all pairs. Table S4 shows the pairwise CCC following the application of ComBat on scans acquired with the LightSpeed Pro 32 model, and resampled to the median IPR value using CWS IM.

Discussion

In this study, the effects of variations in scans’ IPR on the reproducibility of HRFs, the proper methodology of identifying HRFs that are reproducible across different IPRs, and how to properly adjust for these differences before performing radiomics analysis using image interpolation and/or ComBat harmonization were thoroughly investigated. Uniquely, this study evaluates the effects of all the different IMs and the choice of NUIRs on the reproducibility of HRFs. Previous studies usually investigated a single IM with a single NUIR [21,22].

While two batches of scans acquired with the same imaging parameters on two scanner models of the same vendor were used for analysis, the effects of IPR, ComBat, IMs, and

NUIR on the reproducibility of HRFs varied on each of the scanner models. The CCC was calculated pairwise to assess the reproducibility of HRFs when different sets of data were used as batches. Calculating the pairwise CCC between HRF values extracted before resampling the images revealed that the reproducibility of HRFs in our data depended on several factors including, but not limited to, the definition of the HRF, the degree of variation in IPR, and the scanner (hardware) make/model. Addressing the effects of these factors is crucial for performing robust radiomics analysis.

Without performing image preprocessing, the number of reproducible HRFs varied according to the batches being assessed. The aim of this study was to show that different investigated scenarios showed different numbers of reproducible HRFs. Therefore, although 36 HRFs for the Discovery STE scanner (27 HRFs for LightSpeed Pro 32 scanner) were always included in the set of concordant HRFs, it is difficult to conclude that these HRFs are insensitive to spatial resolution on all other scanner models based on our experiments. Yet, our framework guides the methodology of identifying reproducible HRFs according to the data under analysis. As we have shown, the number and type of HRFs is at least sensitive to the scanner model by the same manufacturer. Moreover, we anticipate based on their definition, that certain HRFs (such as histogram-based features) are less sensitive, while others (eg. texture features) are more sensitive to variations in scanning parameters and/or imaging vendors. Generally, scans with more similar original IPRs, and those of integer multiples of IPR showed higher numbers of concordant HRFs before and after resampling. This can be explained by the mechanisms by which a scan is acquired. When all other scanning parameters are fixed, the variations in IPR will result in variations in the number of pixels in 2D, while the other dimensions are pre-served. Therefore, when all other parameters are fixed, the closer the IPR values are, the closer the values of the extracted HRFs.

For the IMs, the number of HRFs that had better/worse concordance after resampling was dependent on the NUIR chosen and scanner model. The window sinc interpolation family performed consistently better on both scanners and NUIRs investigated. In the field of radiology, both NN and linear are known to result in imprecisions [26,35]. A study into the reproducibility of HRFs investigated the performance of B-spline, linear and NN using a single image slice thickness, and concluded that NN is not a favorable method for the reproducibility of HRFs [42]. Our results support these previous reports by showing that NN and linear IMs are not the best candidates for improving the reproducibility of HRFs among scans acquired with different IPRs, and their use led to lower numbers of concordant HRFs in many of the investigated scenarios.

With regard to the selection of NUIR, a common trend of an inverse relationship between the NUIR and the number of concordant HRFs following resampling was

observed. This trend was observed in both scanner models investigated. However, the percentage difference between the concordant HRFs is not significant at the lower end of the NUIR spectrum (Figures 3, 4, S2 and S3; tables S5 and S6). As the best NUIR is expected to be task dependent (for e.g classification of a lesion, predicting response to therapy or overall survival, etc), outcome-based analysis is needed to determine the best NUIR. Yet, as a general rule, the smaller the NUIR, the better the concordance. In addition, while the number of non-highly correlated HRFs was found to be low on both scanner models (9 and 11 HRFs before and after ComBat harmonization, respectively), the exclusion of highly correlated HRFs should be performed based on the effects of the removal of these HRFs on the model performance.

A previous study investigated the effects on HRFs of voxel size resampling using linear interpolation. The authors resampled the scans of a phantom to a single voxel size, which was larger than the largest voxel size in the original scans, and reported that around 20% of the HRFs (N=213) became concordant after resampling [22]. Another study also investigated the effects of voxel size on HRFs of lung cancer patients [21]. The authors resampled all the scans to a single common voxel size using linear interpolation, and reported that resampling does not eliminate all the variations in feature values even when the only variation in scan acquisition and reconstruction parameters was the voxel size, but is favorable to no resampling. Another group investigated the effects of variation in several acquisition and reconstruction parameters on a 13-layer phantom using a different approach, and reported that resampling the scans to isotropic voxels increased the percentage of concordant HRFs from 59.5% to 89.3% [43]. In this study, we found a similar conclusion: the number of previously non-concordant HRFs that became concordant following resampling to the lowest resolution ranged between 1.1% and 22% depending on the IM, and not all HRFs benefit from image resampling.

In contrast to previous studies, we investigated more IMs and harmonization techniques, and propose a guideline on how to carefully approach HRFs reproducibility studies. Furthermore, we found that linear interpolation is not a good candidate for the purpose of improving the reproducibility of HRFs, when compared to other available IMs; and that the performance of an IM is dependent on the original IPR values and the chosen NUIR, as well as the imaging vendor.

When pairwise comparisons were considered, the performance of ComBat harmonization was found to be inferior to that of well-performing IMs, regardless of the NUIR. Moreover, the combination of ComBat and the well-performing IMs did not yield significantly better results compared to solely using the IM. Furthermore, the performance of ComBat varied depending on the batches used. Nevertheless, when the number of concordant HRFs across all pairs was considered, ComBat harmonization was of added value in almost all scenarios. Therefore, ComBat application on HRFs

should follow a reproducibility study (phantom or tissue studies) to assess the impact of ComBat on the reproducibility of HRFs in those settings, and use only the harmonizable HRFs for further radiomics analyses [15], as described in the workflow (Figure 1). The application of ComBat without assessing HRFs' reproducibility as described may result in the inclusion of a high percentage of unreproducible HRFs, or even the loss of some of the HRFs that were originally reproducible, rendering the analysis of these HRFs meaningless. This finding regarding ComBat harmonization is not in line with previous reports, which reported that ComBat successfully removes the batch effects for all HRFs [28,44]. This could be attributed to the differences in the radiomics software and/or the evaluation metrics used. In contrast to previous studies, and as the aim of harmonization is to improve reproducibility but necessarily the performance of generated radiomic models, we opted for the CCC. The CCC provides an accurate description of the reproducibility of HRFs, which is not reflected in neither the distribution of HRFs nor the performance of radiomics models [45]. If radiomic models are to be used clinically, it is expected to be applied to one patient per time. Therefore, the importance has been given in this study to the individual feature values, and not their distributions. HRFs with different values and order rank can share similar distributions, in which case the feature cannot be considered reproducible. In addition, different modeling techniques may yield significantly different results on the same dataset. Hence, the difference in the performance of a radiomic signature before and after harmonization does not necessarily inform about the performance of the harmonization method. Our proposed framework addresses this issue, and guides the selection of reproducible and harmonizable HRFs before developing a radiomic signature, which helps the translation and generalization of results, and ultimately the inclusion of radiomic signatures in clinical practice.

Of note, not all HRFs benefit from resampling all scans to a NUIR, or using ComBat harmonization. Some HRFs lost their concordance following resampling, depending on the IM employed and the chosen NUIR. The combination of IMs and NUIRs affected the HRFs differently on different scanner models. Some HRFs were not found to be concordant on one of the scanner models before or after resampling to any of the available NUIRs using any of the IMs, but were found to be concordant on the other scanner model. Other HRFs were found to be concordant across different scanner models and IPRs. These findings indicate the need for performing reproducibility studies depending on the data under study, and the fact that at this level, we are unable to provide a list of HRFs that can be used regardless of the acquisition and reconstruction parameters and scanner models used. However, it lays down the bases for identifying reproducible HRFs before performing data analysis. In real life scenarios, the variations between the imaging parameters in retrospective cohorts (especially multicentric) are usually not only limited to the IPR. Aside from the scanner/scanning parameters combination variations, some of the effects will be attributed to patient populations. Furthermore, while phantom

studies reflect on the reproducibility of HRFs extracted from anthropomorphic phantoms, HRFs extracted from human tissue are expected to have a wider range of variations, due to the inclusion of biologic factors. This knowledge, combined with our findings, necessitate the critical investigation of the reproducibility of HRFs across the different scanning parameters/scanners before performing any statistical analysis, and future investigations into the effects of differences in acquisition and reconstruction parameters on the re-productibility of HRFs extracted from human tissues, if feasible. Directly performing radiomics analysis on data acquired heterogeneously leads to spurious results, and lacks meaningful interpretation. Henceforth, we reiterate the need for using our proposed robust radiomics analysis framework for addressing differences in IPR. Furthermore, the workflow can be generalized to evaluate other harmonization methods.

Conclusions

The reproducibility of a given HRF, and its harmonizabilty with ComBat are not constants, but depended on the degree of variation in a single reconstruction parameter (the in-plane resolution) of the scans being analyzed. This implies that additional changes in the acquisition and reconstruction parameters could further reduce the number of reproducible and harmonizable HRFs. When scans acquired with different IPR values are to be analyzed, resampling the scans to a unified resolution can significantly improve the reproducibility of HRFs. Interpolation methods (CWS, HWS, BWS, WWS and B-spline) were found to be superior to ComBat harmonization alone in addressing the variations in HRFs attributed to differences in IPR, and the combination of an IM with ComBat following NUIR could increase the number of reproducible HRFs in some scenarios. The application of our proposed framework aids the selection of data- and outcome-specific interpolation and harmonization methods, and is expected to improve the translation and generalizability of radiomics analyses.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: The scanned CCR Phantom, Figure S2: The percentage of concordant features following resampling compared to no resampling with linear trendlines, LightSpeed Pro 32 model, Figure S3: The percentage of concordant features following resampling and ComBat harmonization compared to no resampling with linear trendlines, LightSpeed Pro 32 model, Table S1: Number of pair-wise concordant features with a CCC > 0.9 before resampling, LightSpeed Pro 32 model, Table S2: Number of pair-wise concordant features with a CCC > 0.9 after ComBat, LightSpeed Pro 32 model, Table S3: Number of pair-wise concordant features with a CCC > 0.9 after resampling* using CWS, LightSpeed Pro 32 model, Table S4: Number of pair-wise concordant features with a CCC > 0.9 after ComBat following resampling* using CWS, LightSpeed Pro 32 model,

Table S5: Summary of the number of concordant features before and after resampling, Discovery STE model, Table S6: Summary of the number of concordant features before and after resampling, LightSpeed Pro 32 model, List S1: HRFs with CCC>0.9 across all pairs on Discovery STE model, List S2: HRFs with CCC>0.9 across all pairs on LightSpeed Pro 32 model, List S3: Non-highly correlated HRFs with CCC>0.9 across all pairs on Discovery STE model, List S4: Non-highly correlated HRFs with CCC>0.9 across all pairs on LightSpeed Pro 32 model.

Author Contributions: Conceptualization, Abdalla Ibrahim, Joachim Wildberger, Philippe Lambin and Andrew Maidment; Data curation, Abdalla Ibrahim, Turkey Refae, Sergey Prima-kov, Bruno Barufaldi and Raymond Acciavatti; Formal analysis, Abdalla Ibrahim, Turkey Re-fae, Bruno Barufaldi, Raymond Acciavatti, Renee Granzier and Andrew Maidment; Funding acquisition, Joachim Wildberger; Methodology, Abdalla Ibrahim, Bruno Barufaldi, Raymond Acciavatti, Joachim Wildberger, Philippe Lambin and Andrew Maidment; Project administration, Abdalla Ibrahim, Philippe Lambin and Andrew Maidment; Software, Abdalla Ibrahim; Supervision, Roland Hustinx, Felix Mottaghy, Henry Woodruff, Philippe Lambin and Andrew Maidment; Visualization, Turkey Refae and Sergey Primakov; Writing – original draft, Abdalla Ibrahim, Turkey Refae, Sergey Primakov and Andrew Maidment; Writing – review & editing, Bruno Barufaldi, Raymond Acciavatti, Renee Granzier, Roland Hustinx, Felix Mottaghy, Henry Woodruff, Joachim Wildberger, Philippe Lambin and Andrew Maidment.

Funding: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno). Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2, and Maastricht-Liege Imaging Valley grant, project no. "DEEP-NUCLE".

Conflicts of Interest: Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from radiomics SA, ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals. Dr Lambin has minority shares in the company radiomics SA, Convert pharmaceuticals, MedC2 and LivingMed Biotech, he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed

to Radiomics SA and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (soft-wares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non-licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper. Dr. Woodruff has minority shares in the company OncoRadiomics. Dr. Mottaghy received an advisor fee and reimbursement of travel costs from Oncoradiomics. He reports institutional grants from GE and Nanomab outside the submitted work. Dr. Wildberger reports institutional grants from Agfa, Bard, Bayer, GE, Optimed, Philips, Siemens and personal fees (Speaker's Bureau) from Bayer and Siemens outside the submitted work. The rest of coau-thors declare no competing interest.

References

1. Walsh, S.; de Jong, E.E.C.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.H.M.; et al. Decision Support Systems in Oncology. *JCO Clin Cancer Inform* 2019, 3, 1–9, doi:10.1200/CCI.18.00001.
2. Lambin, P.; Leijenaar, R.T.H.; Deist, T.M.; Peerlings, J.; de Jong, E.E.C.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.H.M.; Even, A.J.G.; Jochems, A.; et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* 2017, 14, 749–762, doi:10.1038/nrclinonc.2017.141.
3. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Refaee, T.; Sanduleanu, S.; Walsh, S.; Morin, O.; et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Semin. Nucl. Med.* 2019, 49, 438–449, doi:10.1053/j.semnuclmed.2019.06.005.
4. Kumar, V.; Gu, Y.; Basu, S.; Berglund, A.; Eschrich, S.A.; Schabath, M.B.; Forster, K.; Aerts, H.J.W.L.; Dekker, A.; Fen-stermacher, D.; et al. Radiomics: the process and the challenges. *Magn. Reson. Imaging* 2012, 30, 1234–1248, doi:10.1016/j.mri.2012.06.010.
5. Miller, A.S.; Blott, B.H.; Hames, T.K. Review of neural network applications in medical imaging and signal pro-cessing. *Med. Biol. Eng. Comput.* 1992, 30, 449–464, doi:10.1007/BF02457822.
6. Kjaer, L.; Ring, P.; Thomsen, C.; Henriksen, O. Texture analysis in quantitative MR imaging. Tissue characterisation of normal brain and intracranial tumours at 1.5 T. *Acta radiol.* 1995, 36, 127–135.
7. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: extracting more information from medical images using advanced feature anal-ysis. *Eur. J. Cancer* 2012, 48, 441–446, doi:10.1016/j.ejca.2011.11.036.
8. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016, 278, 563–577, doi:10.1148/radiol.2015151169.
9. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.H.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.L.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration* 2020, 99, 99–107, doi:10.1159/000505429.
10. Rogers, W.; Thulasi Seetha, S.; Refaee, T.A.G.; Lieverse, R.I.Y.; Granzier, R.W.Y.; Ibrahim, A.; Keek, S.A.; Sanduleanu, S.; Primakov, S.P.; Beuque, M.P.L.; et al. Radiomics: from qualitative to quantitative imaging. *Br. J. Radiol.* 2020, 93, 20190948, doi:10.1259/bjr.20190948.
11. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; Court, L. Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest. Radiol.* 2015, 50, 757–765, doi:10.1097/RLI.0000000000000180.
12. Berenguer, R.; Pastor-Juan, M.D.R.; Canales-Vázquez, J.; Castro-García, M.; Villas, M.V.; Mansilla Legorburo, F.; Sa-bater, S. Radiomics of CT Features May Be Nonreproducible and Redundant: Influence of CT Acquisition Parameters. *Radi-ology* 2018, 288, 407–415, doi:10.1148/radiol.2018172361.
13. Strimbu, K.; Tavel, J.A. What are biomarkers? *Curr. Opin. HIV AIDS* 2010, 5, 463.
14. Davis, A.T.; Palmer, A.L.; Pani, S.; Nisbet, A. Assessment of the variation in CT scanner performance

- (image quality and Hounsfield units) with scan parameters, for image optimisation in radiotherapy treatment planning. *Phys. Med.* 2018, 45, 59–64, doi:10.1016/j.ejmp.2017.11.036.
15. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new frame-work. *Methods* 2020, doi:10.1016/j.ymeth.2020.05.022.
 16. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Wang, J.; Zhang, Z.; Dekker, A.; Lambin, P. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography* 2016, 2, 361–365, doi:10.18383/j.tom.2016.00208.
 17. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of radiomics features in apparent diffusion coefficient maps from a mul-ti-centre test-retest trial. *Sci. Rep.* 2019, 9, 4800, doi:10.1038/s41598-019-41344-5.
 18. Zhovannik, I.; Bussink, J.; Traverso, A.; Shi, Z.; Kalendralis, P.; Wee, L.; Dekker, A.; Fijten, R.; Monshouwer, R. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol* 2019, 19, 33–38, doi:10.1016/j.ctro.2019.07.003.
 19. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* 2018, 102, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.
 20. Papanikolaou, N.; Matos, C.; Koh, D.M. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging* 2020, 20, 33, doi:10.1186/s40644-020-00311-4.
 21. Shafiq-Ul-Hassan, M.; Latifi, K.; Zhang, G.; Ullah, G.; Gillies, R.; Moros, E. Voxel size and gray level normalization of CT radiomic features in lung cancer. *Sci. Rep.* 2018, 8, 10545, doi:10.1038/s41598-018-28895-9.
 22. Shafiq-ul-Hassan, M.; Zhang, G.G.; Latifi, K. Intrinsic dependencies of CT radiomic features on voxel size and num-ber of gray levels. *Medical* 2017.
 23. Thévenaz, P.; Blu, T.; Unser, M. Image interpolation and resampling. of medical imaging, processing and analysis 2000.
 24. Haddad, M.; Porenta, G. Impact of reorientation algorithms on quantitative myocardial SPECT perfusion imaging. *J. Nucl. Med.* 1998, 39, 1864–1869.
 25. Menon, S.; Damian, A.; Hu, S.; Ravi, N.; Rudin, C. PULSE: Self-Supervised Photo Upsampling via Latent Space Explo-ration of Generative Models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; openaccess.thecvf.com, 2020; pp. 2437–2445.
 26. Parker, J.; Kenyon, R.V.; Troxel, D.E. Comparison of interpolating methods for image resampling. *IEEE Trans. Med. Imaging* 1983, 2, 31–39, doi:10.1109/TMI.1983.4307610.
 27. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8, 118–127, doi:10.1093/biostatistics/kxj037.
 28. Orhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of a method to compensate multicenter effects af-fecting CT radiomic features. 2018.
 29. Orhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter

The effects of in-plane spatial resolution on CT-based radiomic features' stability

- Radiomic Studies in PET. *J. Nucl. Med.* 2018, 59, 1321–1328, doi:10.2967/jnumed.117.199935.
30. Mackin, D.; Fave, X.; Zhang, L.; Fried, D.; Yang, J.; Taylor, B.; Rodriguez-Rivera, E.; Dodge, C.; Jones, A.K.; and Court, L. Credence Cartridge Radiomics Phantom CT Scans - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. *Cancer Imaging Archive* 2017.
 31. Clark, K.; Vendt, B.; Smith, K.; Freymann, J.; Kirby, J.; Koppel, P.; Moore, S.; Phillips, S.; Maffitt, D.; Pringle, M.; et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 2013, 26, 1045–1057, doi:10.1007/s10278-013-9622-7.
 32. van Griethuysen, J.J.M.; Fedorov, A.; Parmar, C.; Hosny, A.; Aucoin, N.; Narayan, V.; Beets-Tan, R.G.H.; Fillion-Robin, J.-C.; Pieper, S.; Aerts, H.J.W.L. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017, 77, e104–e107, doi:10.1158/0008-5472.CAN-17-0339.
 33. Lowekamp, B.C.; Chen, D.T.; Ibáñez, L.; Blezek, D. The Design of SimpleITK. *Front. Neuroinform.* 2013, 7, 45, doi:10.3389/fninf.2013.00045.
 34. Hsieh Hou; Andrews, H. Cubic splines for image interpolation and digital filtering. *IEEE Trans. Acoust.* 1978, 26, 508–517, doi:10.1109/TASSP.1978.1163154.
 35. Meijering, E.H.; Niessen, W.J.; Viergever, M.A. Quantitative evaluation of convolution-based methods for medical image interpolation. *Med. Image Anal.* 2001, 5, 111–126, doi:10.1016/s1361-8415(00)00040-2.
 36. Stevenson, M.; Stevenson, M.M.; BiasedUrn, I. Package “epiR.” 2020.
 37. Team, R.C. R language definition. Vienna, Austria: R foundation for statistical computing 2000.
 38. Gandrud, C. *Reproducible Research with R and R Studio*; CRC Press, 2013; ISBN 9781466572843.
 39. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989, 45, 255–268.
 40. McBride, G.B. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA cli-ent report: HAM2005-062 2005, 62.
 41. Zar, J.H. Spearman Rank Correlation. *Encyclopedia of Biostatistics* 2005.
 42. Larue, R.T.H.M.; van Timmeren, J.E.; de Jong, E.E.C.; Feliciani, G.; Leijenaar, R.T.H.; Schreurs, W.M.J.; Sosef, M.N.; Raat, F.H.P.J.; van der Zande, F.H.R.; Das, M.; et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol.* 2017, 56, 1544–1553, doi:10.1080/0284186X.2017.1351624.
 43. Ligeró, M.; Jordi-Ollero, O.; Bernatowicz, K.; Garcia-Ruiz, A.; Delgado-Muñoz, E.; Leiva, D.; Mast, R.; Suarez, C.; Sa-la-Llonch, R.; Calvo, N.; et al. Minimizing acquisition-related radiomics variability by image resampling and batch effect correction to allow for large-scale data analysis. *Eur. Radiol.* 2021, 31, 1460–1470, doi:10.1007/s00330-020-07174-0.
 44. Da-Ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies. *Sci. Rep.* 2020, 10, 10248, doi:10.1038/s41598-020-66110-w.
 45. Vetter, T.R.; Schober, P. Agreement Analysis: What He Said, She Said Versus You Said. *Anesth. Analg.* 2018, 126, 2123–2128, doi:10.1213/ANE.0000000000002924.



76

Chapter 7b

Reply to Orlhac and Buvat on
“Ibrahim *et. al*, The effects of in-plane
spatial resolution on CT-based radiomic
features’ stability with and without
ComBat harmonization *Cancers* 2021,
13, 13081848”

Authors

Abdalla Ibrahim, Turkey Refaee, Sergey P. Primakov, Bruno Barufaldi,
Raymond J. Acciavatti, Renée W.Y Granzier, Roland Hustinx,
Felix M. Mottaghy, Henry C. Woodruff, Joachim E. Wildberger,
Philippe Lambin, Andrew D.A Maidment

Adapted from

Cancers. 2021 Jan;13(12):3080

DOI

10.3390/cancers13123080

We would like to thank Orhac and Buvat [1], for their commentary on our article [2]. Orhac and Buvat present the opinion that we “misused” ComBat harmonization to assess radiomic features in a computed tomography (CT) phantom by evaluating the phantom as a whole. They stated that we must apply ComBat harmonization separately to each layer of the phantom, akin to restricting a radiomics study to either liver or tumour. However, the main aim of our work [2] was not to address a specific radiomics task, but to use CT phantom data to evaluate the robustness of 91 radiomics features to changes in voxel size either alone or with two harmonization methods – interpolation and ComBat.

The application of the ComBat method of Johnson [3] to radiomics, proposed by Fortin *et al.* [4], arose after its initial application to genomics. Johnson sought to harmonize data that were divided into “batches”, “samples”, and “genes”. ComBat “incorporates systematic batch biases common across genes in making adjustments, assuming that phenomena resulting in batch effects often affect many genes in similar ways (i.e. increased expression, higher variability, etc.)”[3]. In the application of ComBat to radiomics, we and Orhac [5] are in agreement that the radiomic features are Johnson’s genes, and that the scans are Johnson’s batches. Thus, the difference comes down to the definition of the sample. Johnson proposed the definition of a sample as being, for example, a patient. By contrast, Orhac and Buvat state “that all measurements grouped in the same batch are *equally* affected by the imaging protocol”[1] (emphasis added), and thus propose that the sample must be a specific texture, for example “liver or tumour” based on the assumption that various textures are affected differently. We believe that this is overly prescriptive. In our usage, the sample is the phantom, which is intended to represent a range of tissue types, because we sought to understand how acquisition differences affect each measure over a range of materials [2]. This is consistent with the use of ComBat by Fortin *et al.* [4].

Consider a simple example – namely that of the first order mean. The phantom in question has 10 layers representing different tissues, including several layers that have a uniform single material. In Figure 1, we show a plot of the paired values of the mean for a single layer and for the whole phantom. By default, all CT scanners use, at a minimum, a two-point calibration of the Hounsfield units (HU), typically performed daily. Nevertheless, CT scans are subject to both stochastic noise arising from the x-radiation and electronic noise in the CT scan, and non-stochastic sources of error in the CT systems such as reconstruction artefacts. However, due to the calibration, the average HU values of a given material in the phantom will be nearly identical in any two scans regardless of pixel size, especially when averaged over large regions. Figure 1a shows the results for a single layer; they are not strongly correlated – nor should they be correlated if the layer represents a single material or a simple admixture of materials. By

contrast, in Figure 1b, we show the results for analysing all phantom layers. As expected, the results are highly correlated since the phantom spans a range of materials. In Figures 2a and 2b, we show the Concordance Correlation Coefficient (CCC) [6] value for the grayscale mean pairwise across the seven scans CCR-2-001 to CCR-2-007 considered in our paper. Note that in analysing a single layer (Figure 2a), we see moderate to no correlation. This arises directly from the physics of imaging objects with limited material differences; the average HU should only vary by stochastic noise and non-stochastic errors. When we analyse all layers (Figure 2b), all scans show high correlation with each other, as expected. Of note, Orhlac and Buvat used ROIs that were smaller in volume than ours, which only serves to increase the stochastic noise, and leads to even more false correlations.

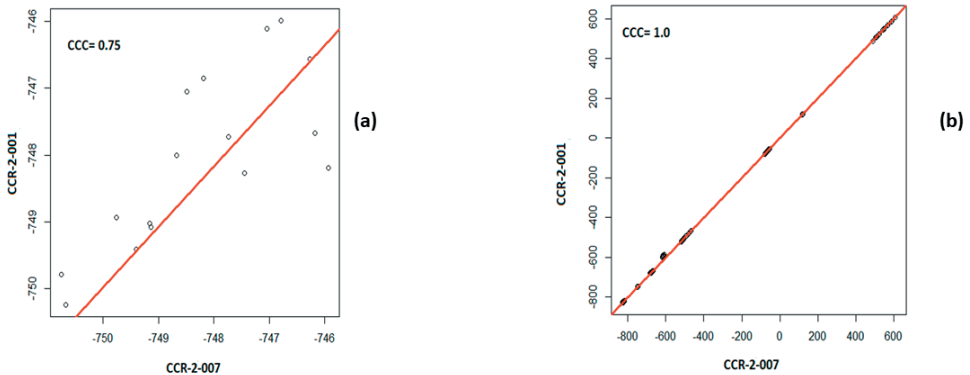


Figure 1: Pairwise plot of the first order mean values with the CCC for (a) a single layer of the phantom (ABS-040), (b) all layers of the phantom, for the scans CCR-2-001 and CCR-2-007.

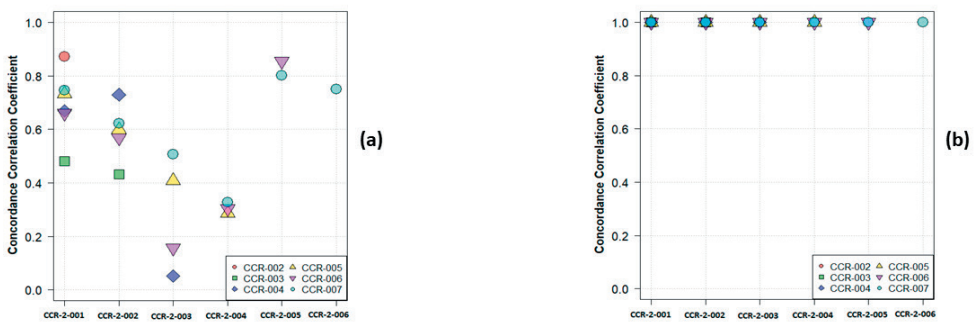


Figure 2: The pairwise CCC values for the first order mean values across 7 scans for (a) a single layer of the phantom (ABS-040), (b) all the layers.

That said, the message of our paper was that ComBat harmonization is not a fix-all. Rather, we argued that one should first apply harmonization steps that directly address physical differences in the acquisition of the images. Fundamental imaging physics dictates that

differences such as voxel size, slice thickness, mAs (dose), and kV can profoundly impact the appearance of images. At least some of these factors, for example voxel size or slice thickness, can be readily harmonized through appropriate and direct image processing, such as resampling. In our paper, we demonstrate that sinc interpolation is superior to pixel replication (nearest neighbour) and other simple interpolation schemes, and that downsampling (harmonization to a coarser resolution) to a common spatial resolution is superior to upsampling. These have simple and obvious physical explanations. However, interpolation to a common pixel size is also not a fix-all. Most importantly, we showed that regardless of the method applied, a reproducibility analysis is required to select reproducible and harmonizable features.

We have also repeated our analysis layer by layer as recommended by Orhlac and Buvat [1] using both parametric and non-parametric ComBat forms, and the results do not change the conclusions of our paper. As suggested, we re-analysed the same scans (CCR-2-001 and CCR-2-007) using 16 cubic volumes of interest (2x2x2 cm³) per layer. In Table 1, we assess the reproducibility of radiomic features before and after ComBat harmonization for each layer separately using the cut-off (CCC>0.9). Indeed, the number of reproducible features before and after ComBat harmonization differ when analysed per layer (Table 1). These results reinforce our original message, that assessing the reproducibility of features with various harmonization methods for each radiomic task is essential. Orhlac and Buvat took the additional step of calculating the CCC for all of the layers after applying ComBat separately for each layer. This presumes a task for which tissue classification or segmentation is applied before ComBat harmonization. This is task dependent; for example, Verma *et al.* [7], considered analysis of grey matter and white matter both separately and jointly, but found no difference in performance.

Table I. The number of reproducible radiomic features for the different phantom layers between scan CCR-2-001 and CCR-2-007.

Phantom layer	Number (%) before ComBat harmonization	Number (%) after ComBat harmonization	
		Parametric	Non-parametric
ABS-020	0 (0.0%)	3 (3.3%)	3 (3.3%)
ABS-030	0 (0.0%)	1 (1.1%)	0 (0.0%)
ABS-040	0 (0.0%)	3 (3.3%)	3 (3.3%)
ABS-050	3 (3.3%)	14 (15.4%)	9 (9.9%)
Wood	27 (29.7%)	38 (41.2%)	36 (39.6%)
Rubber	2 (2.2%)	36 (39.6%)	31 (36.3%)
Dense Cork	6 (6.6%)	26 (28.6%)	24 (26.4%)
Acrylic	6 (6.6%)	32 (35.2%)	32 (35.2%)
Cork	7 (7.7%)	42 (46.2%)	35 (38.5%)
Resin	22 (24.2%)	44 (48.4%)	41 (45.1%)



Orlhac and Buvat also state that the definition of the design matrix of covariates affects the outcome of ComBat [1]. We agree. The aim of the design matrix of the biologic covariates in ComBat is to preserve biologic information while harmonizing the features [3,8,9]. However, as we have stated [2], we performed this study to examine the impact of pixel interpolation on radiomic features in a phantom, and no biologic covariates were appropriate for our study. We clearly state in the discussion that anthropomorphic phantom scans provide some evidence into the reproducibility of features, but that they cannot completely represent features extracted from human images, and human or cadaveric reproducibility studies are encouraged when ethical.

In summary, we disagree with the statement of Orhac and Buvat that we “misused” ComBat [1]. First, their method of application to a specific material (or in the case of the phantom, a single layer) will not express the full impact of the underlying imaging physics, which we were trying to elicit in our study. Second, by choosing ROI sizes that are sensitive to stochastic noise, Orhac and Buvat run the risk of overfitting image noise and producing false correlations. Third, Orhac and Buvat suppose that all radiomic tasks require the same definition of the “sample” be used. For this, we fundamentally disagree; the choice of sample depends upon the task. We do agree with Orhac and Buvat that the design matrix can affect the outcome of ComBat. Finally, it is worth noting that as described in our paper, we used Pyradiomics version 2.1.2 which has 91 features, and Orhac used Pyradiomics version 3.0.0 which has 93 features; this accounts for the difference in features between our work and Orhac [1].

Thus, the message of our study [2] remains unchanged: 1) image interpolation is a useful harmonization method to address variations in pixel spacing; 2) ComBat harmonization was of added value in almost all scenarios; 3) the effects of interpolation and ComBat on the reproducibility of radiomic features is dependent on the data being analysed; 4) neither interpolation nor ComBat is a *fix-all*; and 5) regardless of the harmonization method applied, study data should be analysed to identify *reproducible features* and used to help interpret and generalize radiomic models developed with these features.

Supplementary Materials:

Author Contributions: Conceptualization, Abdalla Ibrahim, Joachim Wildberger, Philippe Lambin and Andrew Maidment; Data curation, Abdalla Ibrahim, Turkey Refaee, Sergey Primakov, Bruno Barufaldi and Raymond J. Acciavatti; Formal analysis, Abdalla Ibrahim, Turkey Refaee, Bruno Barufaldi, Raymond J. Acciavatti, Renee Granzier and Andrew Maidment; Funding acquisition, Joachim Wildberger; Methodology, Abdalla Ibrahim, Bruno Barufaldi, Raymond J. Acciavatti, Joachim Wildberger, Philippe Lambin and Andrew Maidment; Project administration, Abdalla Ibrahim, Henry C. Woodruff,

Philippe Lambin and Andrew Maidment; Software, Abdalla Ibrahim and Turkey Refaee; Supervision, Roland Hustinx, Felix Mottaghy, Philippe Lambin and Andrew Maidment; Visualization, Turkey Refaee and Sergey Primakov; Writing – original draft, Abdalla Ibrahim, Turkey Refaee, Sergey Primakov and Andrew Maidment; Writing – review & editing, Bruno Barufaldi, Raymond J. Acciavatti, Renee Granzier, Roland Hustinx, Felix Mottaghy, Henry C. Woodruff, Joachim Wildberger, Philippe Lambin and Andrew Maidment.

Funding: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno). Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2, and Maastricht-Liege Imaging Valley grant, project no. "DEEP-NUCLE".

Conflicts of Interest: Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from radiomics SA, ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals. Dr Lambin has minority shares in the company radiomics SA, Convert pharmaceuticals, MedC2 and LivingMed Biotech, he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non-licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper. Dr. Woodruff has minority shares in the company OncoRadiomics. Dr. Mottaghy received an advisor fee and reimbursement of travel costs from Oncoradiomics. He reports institutional grants from GE and Nanomab outside the submitted work. Dr. Wildberger reports institutional grants from Agfa, Bard, Bayer, GE, Optimed, Philips, Siemens and personal fees (Speaker's Bureau) from Bayer and Siemens outside the submitted work. The rest of coauthors declare no competing interest.

References

1. Orlhac, F.; Buvat, I. Comment on Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848. *Cancers* **2021**.
2. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848.
3. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **2007**, *8*, 118-127.
4. Fortin, J.-P.; Parker, D.; Tunç, B.; Watanabe, T.; Elliott, M.A.; Ruparel, K.; Roalf, D.R.; Satterthwaite, T.D.; Gur, R.C.; Gur, R.E. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage* **2017**, *161*, 149-170.
5. Orlhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of a method to compensate multicenter effects affecting CT Radiomics. *Radiology* **2019**, 182023.
6. Lawrence, I.; Lin, K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **1989**, 255-268.
7. Verma, R.; Swanson, R.L.; Parker, D.; Ismail, A.A.O.; Shinohara, R.T.; Alappatt, J.A.; Doshi, J.; Davatzikos, C.; Gallaway, M.; Duda, D. Neuroimaging findings in US Government personnel with possible exposure to directional phenomena in Havana, Cuba. *Jama* **2019**, *322*, 336-347.
8. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods* **2020**.
9. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLOS ONE* **2021**, *16*, e0251147, doi:10.1371/journal.pone.0251147.

A large, stylized white number 8 is centered on a blue watercolor splash background. The splash is composed of various shades of blue, from light to dark, with some darker spots and a textured, organic appearance. The number 8 is a simple, clean, sans-serif font, standing out prominently against the blue background.

8

Chapter 8

Reproducibility of CT-based Hepatocellular carcinoma radio-mic features across different contrast imaging phases: A proof of concept on SORAMIC trial data

Authors

Abdalla Ibrahim, Yousif Widaatalla, Turkey Refaee, Sergey Primakov,
Razvan L. Miclea, Osman Öcal, Matthias P. Fabritius, Michael Ingrisich, Jens Ricke,
Roland Hustinx, Felix M. Mottaghy, Henry C. Woodruff, Max Seidensticker
and Philippe Lambin

Adapted from

Cancers. 2021 Jan;13(18):4638

DOI

10.3390/cancers13184638

Abstract

Handcrafted radiomic features (HRFs) are quantitative imaging features extracted from regions of interest on medical images, which can be correlated with clinical outcomes and biologic characteristics. While HRFs have been used to train predictive and prognostic models, their reproducibility has been reported to be affected by variations in scan acquisition and reconstruction parameters, even within the same imaging vendor. In this work, we evaluated the reproducibility of HRFs across the arterial and portal venous phases of contrast enhanced computed tomography images depicting hepatocellular carcinomas, as well as the potential of ComBat harmonization to correct for this difference. ComBat harmonization is a method based on Bayesian estimates that was developed for gene expression arrays, and has been investigated as a potential method for harmonizing HRFs. Our results show that the majority of HRFs are not reproducible between the arterial and portal venous imaging phases, yet a number of HRFs could be used interchangeably between those phases. Furthermore, ComBat harmonization increased the number of reproducible HRFs across both phases by 1%. Our results guide the pooling of arterial and venous phases from different patients in an effort to increase cohort size, as well as joint analysis of the phases.

Keywords

Hepatocellular carcinoma; CT radiomics; domain translation; reproducibility.

Introduction

The recent decades witnessed vast advances in computational power, artificial intelligence, and medical imaging techniques [1], which provided a unique opportunity for transforming the abundant amounts of medical imaging into mineable quantitative data. The concept acquired much scientific attention recently, and a branch of medical imaging analysis -known as handcrafted radiomics- emerged as a result [2]. Handcrafted Radiomic features (HRFs) are quantitative features extracted with high throughput from medical imaging, with its varying modalities. The hypothesis is that medical images carry more data than can be seen by trained human eyes, and that these data can be decoded using the HRFs, i.e correlations between HRFs and underlying biology could potentially exist [3]. Since the introduction of the field, many studies reported on the potential of radiomic signatures to predict clinical endpoints, the majority of which were performed on computed tomography (CT) [4–7], magnetic resonance (MR) [8–10], and positron emission tomography (PET) scans [11,12].

Hepatocellular carcinoma (HCC) is the most common primary liver cancer, the fifth most common malignancy worldwide, and a leading cause of cancer-related mortality [13]. Different diagnostic approaches and treatment modalities are used clinically depending on the characteristics of the patient and the progression of the disease [14,15]. Contrast-enhanced computed tomography (CE-CT) scans are considered one of the main diagnostic tools for HCC. CE-CT can be acquired at different times following the injection of the contrast agent to acquire arterial, venous or late phase scans. Each phase shows specific characteristics for HCC lesions. However, there is still a clinical need for reliable non-invasive tools that could aid diagnosing and devising individualized treatment plans for HCC patients. Several studies investigated and reported on the potential of HRFs to aid clinical decision making in HCC patients [16–19].

While numerous studies have reported on the potential of HRFs in aiding clinical decision making on HCC and other diseases, several hurdles hindering the clinical translation of radiomic signatures to clinical decision support systems have been identified. These hurdles include the reproducibility of HRFs in test-retest studies, their sensitivity to variations in acquisition and reconstruction parameters of the scans, inter-observer variability, and the need for big data [20–26]. However, the need for big data in radiomics analysis necessitates the exploration of methods for combining and comparing retrospective medical imaging databases.

A number of studies tried to address the issue of reproducibility of HRFs using ComBat harmonization [27–30]. ComBat harmonization is a method that was developed to remove the batch effects in gene expression arrays [31]. The studies that investigated the

application of ComBat in radiomics analyses reported on the improvement in performance metrics of developed radiomic signatures after the application of ComBat compared to before, and recommended the use of the method. Other studies that investigated the reproducibility of HRFs on phantom datasets acquired with different settings [32], or with a single parameter difference [33], and reported that the performance of ComBat is dependent on the data under study and recommended a framework to assess the reproducibility of HRFs. Yet to date, no study reported on the agreement in HRFs across different phases or the potential of ComBat to remove the effects of different imaging phases from HRFs, which could allow the proper combination of phases in a single analysis, or the interchangeability of HRFs across phases to allow the use of different imaging scans per patient. Furthermore, no study performed a reproducibility analysis for HRFs following ComBat harmonization on patients' scans acquired with a single parameter difference.

We hypothesize that the time of acquisition after the injection of the contrast agents adds another level of complexity to be accounted for in the radiomics analysis, as HRFs might be affected by the appearance of contrast, due to the variations in the distribution of the contrast within the lesions. As a proof of concept, we investigate the sensitivity of HRFs extracted from CE-CT scans depicting HCC acquired during the arterial and portal venous phases, when all other acquisition and reconstruction parameters were fixed. Furthermore, we investigate the potential of the ComBat harmonization for domain translation of the HRFs extracted from these scans. Ultimately, we aim to (i) guide the identification of HRFs that can be used interchangeably between arterial and venous phase scans, which could increase the number of scans that can be included in a CE-CT based radiomics study; and (ii) identify the features that can be used in studies analyzing both phases simultaneously to maximize the information extracted from ROIs

Materials and Methods

Patients and Imaging data

The imaging data were originally collected for the European multicenter clinical trial (SORAMIC) [34]. Imaging data for 424 patients diagnosed with HCC (using cyto-histological criteria, radiologic criteria, or a combination of both) were obtained for the SORAMIC trial, of which 338 scans were available for analysis in this study. Scans that contained artifacts were considered of poor quality (n=48). From the available 338 patients with both arterial and portal venous scans available, patients with scans that had any difference in the acquisition or reconstruction parameters, or lacked segmentations re-viewed by an expert, were excluded. A total of 61 patients with 104 distinct lesions were finally included in this study (Figure 1). Scans included were acquired from different hospitals,

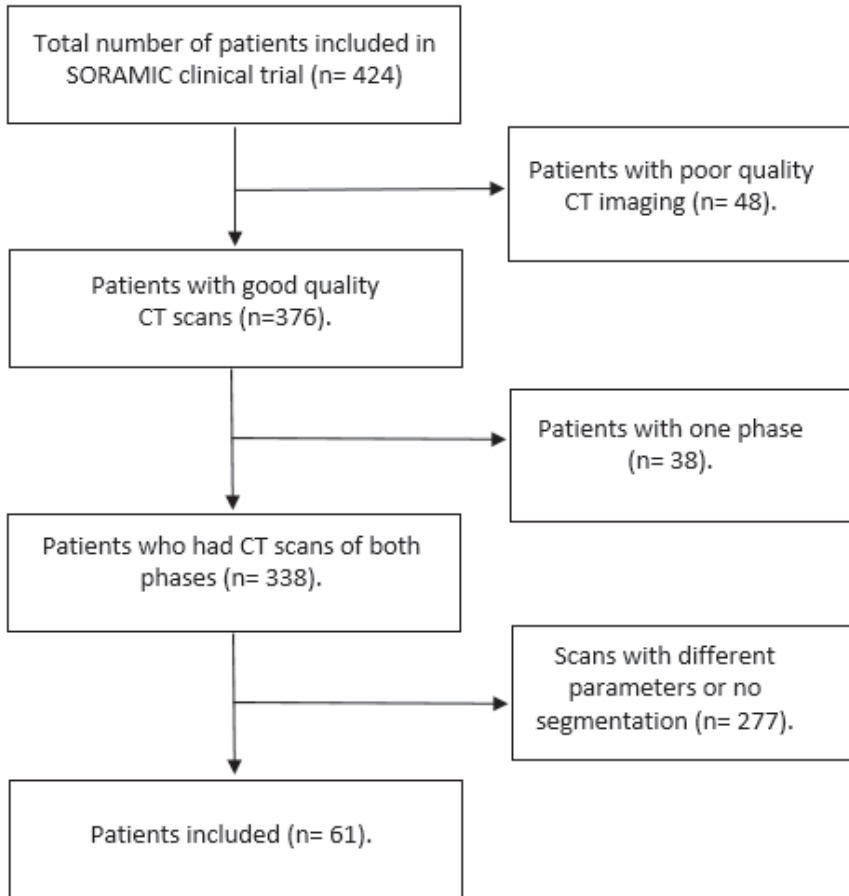


Figure I: A flowchart showing the patients selection process.

using different vendors and protocols. In total, 9 scanner models from 4 different imaging vendors, and a range of scanning parameters, were included, as shown in Table 1. The imaging analysis was approved by the University of Magdeburg institutional review board (IRB00006099, EudraCT no 2009-012576-27), and informed consent was obtained from all included patients. All methods were carried out in accordance with the relevant guidelines and regulations [35].

Table I: Acquisition and reconstruction parameters for the imaging dataset.

Manufacturer	Scanner model	X-Ray Tube Current (kV)	Exposure (mAs)	Convolution kernels	Slice thickness (mm)	Pixel spacing (mm ²)
TOSHIBA	Aquilion	50 - 360	2-300	FC13	1-5	0.39x0.39 - 0.98x0.98
	Aquilion PRIME					
Philips	Brilliance 64			B		
GE	Discovery CT750 HD			STANDARD		
	Optima CT660					
SIEMENS	Sensation 16			B31f		
	SOMATOM Definition AS					
	SOMATOM Definition Flash			I30f, I40f		
	SOMATOM Force			Br40d		

Segmentation and HRFs extraction

The scans of a single patient were co-registered. The region of interest (ROI) was segmented on each scan while viewing both phases simultaneously and saved to both scans (Fig 2). The segmentations were performed using MIM software (MIM Software Inc., Cleveland, OH) by a medical doctor (Y.W) with 2 years of experience in image segmentation, and revised by a radiologist (R.M.) with 15 years of experience in medical radiology.

HRFs were extracted from these ROIs using the software RadiomiX Discovery Toolbox (version, October 2019; <https://www.radiomics.bio>), which calculates HRFs compliant with the Imaging Biomarkers Standardization Initiative (IBSI) [36], in addition to others. Image intensities were binned with a binwidth of 25 Hounsfield Units (HUs) in order to reduce noise levels and to reduce texture matrix sizes, and therewith computation power, with no resampling or further preprocessing of the images. The description of the extracted HRFs was published previously [24].

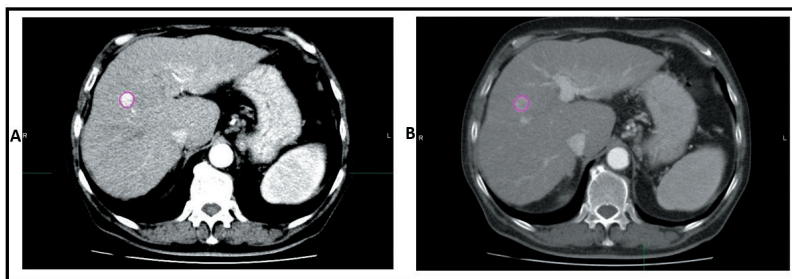


Figure 2: An example of ROI segmented in (A) the arterial phase and (B) portal venous phase.

ComBat Harmonization

ComBat method employs empirical Bayes to estimate the effects of assigned batches on the data being harmonized. For HRFs, ComBat assumes that a feature value can be approximated by the equation:

$$Y_{ij} = \alpha + \beta X_{ij} + \gamma_i + \delta_i \varepsilon_{ij} \quad (1)$$

where α is the average value for HRF Y_{ij} for ROI j on scanner i ; X is a design matrix of the biologic covariates that are known to affect the value of HRFs; β is the vector of regression coefficients corresponding to each biologic covariate; γ_i is the additive effect of scanner i on HRFs, δ_i is the multiplicative scanner effect, and ε_{ij} is an error term, presupposed to be normally distributed with zero mean. Based on the values estimated, ComBat performs feature transformation as given by the formula:

$$Y_{ij}^{ComBat} = \frac{(Y_{ij} - \hat{\alpha} - \hat{\beta} X_{ij} - \gamma_i^*)}{\delta_i^*} + \hat{\alpha} + \hat{\beta} X_{ij} \quad (2)$$

where $\hat{\alpha}$ and $\hat{\beta}$ are estimators of the parameters α and β , respectively; and γ_i^* and δ_i^* are the empirical Bayes estimates for the parameters γ_i and δ_i , respectively.

Statistical Analysis

All statistical analyses were performed using R language [37] on RStudio (V 3.6.3) [38]. To determine the reproducibility of HRFs, the concordance correlation coefficient (CCC) between the HRFs values across the two phases was calculated [39], using epiR package [40]. The CCC measures how concordant are the values of a given HRF and the rank of each data point relative to the rest in each batch. HRFs with $CCC > 0.9$ were considered reproducible and could be interchangeably used between the arterial and venous phase CT scans.

To assess the performance of ComBat, shape features and HRFs with (near) zero variance (HRFs that have the same value in 95% or more of the observations) were removed. The phase of the scan was assigned as the batch for ComBat harmonization. The CCC was calculated after ComBat application and the cutoff of $CCC > 0.9$ was applied to select the concordant HRFs. The correlation of concordant features with volume was assessed using Pearson correlation. Features that had a correlation coefficient > 0.85 were considered highly correlated. The analysis code used in this study can be found on: (<https://github.com/AbdallaIbrahim/The-reproducibility-and-ComBat-ability-of-Radiomic-features>).

Results

Patient characteristics

The patients included (n=61) had a median age of 66 years, mainly male (n=50, 81.9%), with cirrhotic livers (n=56, 91.8%), and a minority (n=11, 18.1%) had portal vein invasion. For more patient characteristics see Table 2.

Table 2: Patient characteristics.

Characteristic	N=61
Gender, male (%)	50 (81.9%)
Age, median (range)	66 (48-81)
Cirrhosis, yes (%)	56 (91.8%)
Child-Pugh grade	
A	56 (91.8%)
B	5 (8.2%)
Diameter of largest lesion, in mm, median (range)	37 (10-220)
Portal vein invasion, yes (%)	11 (18.1%)
Extrahepatic disease yes (%)	7 (11.4%)
BCLC staging	
A	22 (36.1%)
B	22 (36.1%)
C	17 (27.8%)
ECOG performance	
0	58 (95.1%)
1	3 (4.9%)

* Barcelona Clinic Liver cancer (BCLC) staging

** European Cooperative Oncology Group (ECOG) performance

Extracted HRFs

A total of 167 original HRFs were extracted from each of the available 104 ROIs. These HRFs are divided into 11 feature families: Fractal (n=3), Gray Level Co-occurrence Matrix (GLCM; n= 26), Gray Level Distance Zone Matrix (GLDZM; n=16), Gray Level Run Length Matrix (GLRLM; n=15), Gray Level Size Zone Matrix (GLSZM, n=16), Intensity Histogram (IH; n=25), Local Intensity (LocInt, n=2), Neighbouring Gray Level De-pendence Matrix (NGLDM; n=17), Neighbouring Gray Tone Difference Matrix (NGTDM, n=5), Shape (n=23), and Statistics (Stats, n=19).

The effects of differences in imaging phase on the reproducibility of HRFs

Out of the 167 extracted HRFs, 42 (25%) were reproducible (had a CCC>0.9) across both phases (Figure 3a, shape features were not included to ease the comparison between figures). These HRFs were divided into shape (n=22), NGTDM (n=1), NGLDM (n=4),

Reproducibility of CT-based Hepatocellular carcinoma radio-mic features

IH (n=2), GLSZM (n=4), GLRLM (n=2) and GLDZM (n=7). The remaining HRFs had a CCC ranging from -0.07 and 0.85, with a median of 0.39.

Of the concordant 22 shape features, 8 features were highly correlated with volume ($R > 0.85$), in addition to 1 feature from the NGLDM group (NGLDM_DN) and 2 features from the GLRLM group (GLRLM_RLN and GLRLM_GLN). The remaining features (31, 73.8%) had a correlation coefficient < 0.85 .



Figure 3: (a) The CCC values for the different HRFs before ComBat harmonization; (b) The CCC values for the different HRFs after ComBat harmonization.

The effects of ComBat on the reproducibility of HRFs

The application of ComBat harmonization to remove the batch effects attributed to the difference in time between contrast injection and scan acquisition resulted in a total of 44 (26.1%) reproducible HRFs, i.e 2 extra HRFs became concordant following the application of ComBat: Stats_energy and GLDZM_HILDE (Fig 3b). The remaining 20 HRFs had a CCC>0.9 before and after ComBat harmonization, in addition to the shape features (n=22). The CCC of stats_energy increased from 0.8 to 0.95 following ComBat harmonization, and that of GLDZM_HILDE increased from 0.34 to 0.93.

The impact of ComBat on the CCC values had a wide range; 6 HRFs had an increment in CCC between 0.5 and 0.6; 42 HRFs had an increment in CCC between 0.1 and 0.49; 87 HRFs had an increment between 0 and 0.09; and 33 HRFs had a decrement in CCC between -0.001 and -0.06. Following ComBat harmonization, the number of highly correlated features with volume increased by one feature (Stats_energy). The concordant features before domain translation maintained their correlation with the volume.

Discussion

In this study, we investigated the reproducibility of HCC CT-based HRFs across the arterial and portal venous imaging phases when all other scanning parameters were fixed, and whether ComBat harmonization improves the reproducibility of HRFs in such a scenario. Uniquely, this is the first manuscript to investigate the potential of ComBat to remove batch effects attributed to the differences in imaging phase, and on patient data with a single parameter difference between the compared/harmonized scans. Our results show that the majority of HRFs were significantly affected by the difference in imaging phases, and only a quarter of the total extracted number of HRFs were reproducible across both phases. Moreover, ComBat harmonization did not successfully harmonize the majority of HRFs, even though the differences between the batches compared were limited to the variations in imaging phase.

HRFs are calculated using mathematical formulas applied on the array of values representing the medical image [41]. Changes in the value of units in this array are expected to have an impact on the value calculated by the same formula. Therefore, changes in the scanning parameters are expected to affect the reproducibility of different HRFs variably. Aside from HRFs that are not reproducible in test-retest studies, the sensitivity of the remaining HRFs to the imaging phase can be justified by the increased radio-opaqueness and the resulting perfusion patterns of contrast within the ROI, and thus, changes in the image array values based on which the HRFs are calculated. As expected, statistics and intensity histogram features, which are simple HRFs based on

a single voxel value (e.g. minimum or maximum intensity value) or the description of their distribution (e.g. mean or median intensity value), were found to be the most significantly affected families. On the other hand, also according to expectations, HRFs that do not depend on the intensity values, but the shape of the segmentation (shape features), were found to be reproducible across both phases, with the exception of the shape feature centroid distance, which is based on the distribution of intensity values around the geometric center of the ROI. The copying of segmentations and the inclusion of scans that were acquired identically in both phases allowed isolating the effects of differences imaging phases on HRFs. However, in scenarios where acquisition and/or reconstruction parameters, or the segmentation of the ROI changes, the reproducibility of HRFs is expected to be further impacted. This is also in line with what reported in a study that investigated the reproducibility of liver parenchyma and tumors HRFs extracted from two contrast enhanced scans (one phase) taken within a 14 days interval [42]. Therefore, the reproducibility analysis based on the data under study should be an integral part of each radiomics study.

Our study sheds the light on the methodology of combining HRFs from different modalities, either for the purpose of combining different phases/modalities per patient, or the combination of different phases for different patients. For merging different modalities per patient, we show that a number of HRFs is reproducible across the phases. Therefore, models that try to combine different imaging phases per patient are recommended to define which reproducible (test-retest) HRFs vary across the available phases, and pre-select those for further analysis. Another implication of our findings is allowing the combination of different imaging phases per patient (e.g due to the lack of data), when only the reproducible HRFs across phases are extracted and compared between the different patients, regardless of the available imaging phase for each patient. This approach can significantly increase the number of data points in retrospective radiomics studies.

The correlation of radiomic features with the volume of the ROI has been considered one of the major points to be assessed in radiomics analysis, since some of the features were reported previously to be surrogates of volume [43]. In our analysis, we observed that the majority of the features identified as concordant (or domain-translatable with ComBat) between the arterial and venous CT scans was considerable, most of which were shape features. However, the majority of features were not found to be highly correlated with volume, which means that these features can decode additional information about the ROIs being investigated.

The number of features that had a CCC value higher than 0.9 was slightly higher after the application of ComBat on the HRFs extracted from the arterial and portal

venous phases. ComBat successfully harmonized two additional HRFs compared to the number of concordant HRFs before domain translation. The majority of HRFs were not concordant across the phases even after the application of ComBat harmonization. The differences in ComBat performance per HRF (and feature families) are also expected, as in contrast to gene expression arrays, HRFs have different levels of complexity and are not expected to be uniformly affected by the batch defined for domain translation. The variant performance of ComBat on HRFs could be explained by the differences in the complexity of HRFs, compared to gene expression arrays [21].

The findings are in line with the reproducibility studies that assessed the performance of ComBat on phantom scans, which reported that ComBat harmonization does not successfully harmonize all HRFs, and that its performance is dependent on the variations between the batches [32,33]. As a consequence, we recommend that the application of ComBat harmonization on HRFs follows a reproducibility analysis with reference values to assess its performance, as it is expected to vary with the variations in the dataset batches being harmonized [21]. Other deep learning based harmonization methods that have been recently investigated [44–47] might be more suitable for domain translation of images acquired in different phases. However, this is yet to be investigated.

While this study provides a proof of concept for the combination/replacement of different imaging phases, we speculate that the set of reproducible HRFs identified in this study is limited to HCC lesions extracted from scans acquired similarly to our dataset. Furthermore, the changes in reconstruction parameters (and sometimes acquisition parameters) between the two imaging phases in clinical routine significantly lowered the number of available scans to perform this analysis. Lastly, the reproducibility of the identified HRFs has to be investigated across different acquisition and reconstruction parameters. However, due to the lack of data, this was not performed. Nevertheless, this study serves as a guide for selecting and/or harmonizing the reproducible HRFs in future radiomic studies that utilize contrast enhanced imaging.

Conclusions

The majority of HRFs are significantly affected by changes in the imaging phase of the scan. Studies that investigate the potential of combining HRFs from different imaging phases or modalities must investigate the reproducibility and interoperability of the HRFs across the investigated phases for the lesions of interest. Furthermore, a number of HRFs can be interchangeably used between the arterial and portal venous phases, and these can be used to increase data points in retrospective imaging studies. ComBat harmonization increased the number of comparable CT based HRFs across the arterial and portal venous imaging phases for HCC lesions by 1% in our dataset.

Supplementary Materials:

Author Contributions: Conceptualization, Abdalla Ibrahim, Yousif Widaatalla, Turkey Refae and Philippe Lambin; Data curation, Abdalla Ibrahim, Yousif Widaatalla, Turkey Refae, Razvan Miclea, Osman Öcal, Michael Ingrisich, Jens Ricke and Max Seidensticker; Formal analysis, Abdalla Ibrahim and Turkey Refae; Methodology, Abdalla Ibrahim, Yousif Widaatalla, Turkey Refae, Max Seidensticker and Philippe Lambin; Project administration, Abdalla Ibrahim, Max Seidensticker and Philippe Lambin; Software, Abdalla Ibrahim and Turkey Refae; Supervision, Roland Hustinx, Felix M. Mottaghy, Henry Woodruff, Max Seidensticker and Philippe Lambin; Visualization, Turkey Refae and Sergey Primakov; Writing – original draft, Abdalla Ibrahim, Yousif Widaatalla, Turkey Refae and Sergey Primakov; Writing – review & editing, Abdalla Ibrahim, Yousif Widaatalla, Turkey Refae, Sergey Primakov, Razvan Miclea, Osman Öcal, Michael Ingrisich, Jens Ricke, Roland Hustinx, Felix M. Mottaghy, Henry Woodruff, Max Seidensticker and Philippe Lambin..

Funding: Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno). Authors also acknowledge financial support from the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295). Authors further acknowledge financial support by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2, and Maastricht-Liege Imaging Valley grant, project no. "DEEP-NUCLE".

Institutional Review Board Statement: The imaging analysis was approved by the University of Magdeburg institutional review board (IRB00006099, EudraCT no 2009-012576-27).

Informed Consent Statement: Informed consent was obtained from all included patients.

Data Availability Statement: The data is privately owned by the trial coordinators.

Conflicts of Interest: Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from radiomics SA, ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals.

Chapter 8

Dr. Lambin has minority shares in the company radiomics SA, Convert pharmaceuticals, MedC2, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (soft-wares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non-licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper. Dr. Woodruff has minority shares in the company OncoRadiomics. Dr. Mottaghy received an advisor fee and reimbursement of travel costs from Oncoradiomics. He reports institutional grants from GE and Nanomab outside the submitted work. The rest of co-authors declare no competing interest.

References

1. Walsh, S.; de Jong, E.E.C.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T.H.M.; et al. Decision Support Systems in Oncology. *JCO Clin Cancer Inform* 2019, 3, 1–9, doi:10.1200/CCI.18.00001.
2. Lambin, P.; Rios-Velazquez, E.; Leijenaar, R.; Carvalho, S.; van Stiphout, R.G.P.M.; Granton, P.; Zegers, C.M.L.; Gillies, R.; Boellard, R.; Dekker, A.; et al. Radiomics: Extracting More Information from Medical Images Using Advanced Feature Analysis. *Eur. J. Cancer* 2012, 48, 441–446, doi:10.1016/j.ejca.2011.11.036.
3. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology* 2016, 278, 563–577, doi:10.1148/radiol.2015151169.
4. Refaee, T.; Wu, G.; Ibrahim, A.; Halilaj, I.; Leijenaar, R.T.H.; Rogers, W.; Gietema, H.A.; Hendriks, L.E.L.; Lambin, P.; Woodruff, H.C. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration* 2020, 99, 99–107, doi:10.1159/000505429.
5. Aerts, H.J.W.L. The Potential of Radiomic-Based Phenotyping in Precision Medicine: A Review. *JAMA Oncol* 2016, 2, 1636–1642, doi:10.1001/jamaoncol.2016.2631.
6. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Reymen, B.; Oberije, C.; Monshouwer, R.; Bussink, J.; Brink, C.; Hansen, O.; Lambin, P. Survival Prediction of Non-Small Cell Lung Cancer Patients Using Radiomics Analyses of Cone-Beam CT Images. *Radiother. Oncol.* 2017, 123, 363–369, doi:10.1016/j.radonc.2017.04.016.
7. Panth, K.M.; Leijenaar, R.T.H.; Carvalho, S.; Lieuwes, N.G.; Yaromina, A.; Dubois, L.; Lambin, P. Is There a Causal Relationship between Genetic Changes and Radiomics-Based Image Features? An in Vivo Preclinical Experiment with Doxycycline Inducible GADD34 Tumor Cells. *Radiother. Oncol.* 2015, 116, 462–466, doi:10.1016/j.radonc.2015.06.013.
8. Jethanandani, A.; Lin, T.A.; Volpe, S.; Elhalawani, H.; Mohamed, A.S.R.; Yang, P.; Fuller, C.D. Exploring Applications of Radiomics in Magnetic Resonance Imaging of Head and Neck Cancer: A Systematic Review. *Front. Oncol.* 2018, 8, 131, doi:10.3389/fonc.2018.00131.
9. Ursprung, S.; Beer, L.; Bruining, A.; Woitek, R.; Stewart, G.D.; Gallagher, F.A.; Sala, E. Radiomics of Computed Tomography and Magnetic Resonance Imaging in Renal Cell Carcinoma—a Systematic Review and Meta-Analysis. *Eur. Radiol.* 2020, 30, 3558–3566, doi:10.1007/s00330-020-06666-3.
10. Samiei, S.; Granzier, R.W.Y.; Ibrahim, A.; Primakov, S.; Lobbes, M.B.I.; Beets-Tan, R.G.H.; van Nijntten, T.J.A.; Engelen, S.M.E.; Woodruff, H.C.; Smidt, M.L. Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer. *Cancers* 2021, 13, doi:10.3390/cancers13040757.
11. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Refaee, T.; Sanduleanu, S.; Walsh, S.; Morin, O.; et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Semin. Nucl. Med.* 2019, 49, 438–449, doi:10.1053/j.semnuclmed.2019.06.005.
12. Lovinfosse, P.; Visvikis, D.; Hustinx, R.; Hatt, M. FDG PET Radiomics: A Review of the Methodological Aspects. *Clinical and Translational Imaging* 2018, 6, 379–391, doi:10.1007/

- s40336-018-0292-9.
13. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* 2021, doi:10.3322/caac.21660.
 14. Aubé, C.; Oberti, F.; Lonjon, J.; Pageaux, G.; Seror, O.; N’Kontchou, G.; Rode, A.; Radenne, S.; Cassinotto, C.; Vergniol, J.; et al. EASL and AASLD Recommendations for the Diagnosis of HCC to the Test of Daily Practice. *Liver Int.* 2017, 37, 1515–1525, doi:10.1111/liv.13429.
 15. Finn, R.S.; Qin, S.; Ikeda, M.; Galle, P.R.; Ducreux, M.; Kim, T.-Y.; Kudo, M.; Breder, V.; Merle, P.; Kaseb, A.O.; et al. Atezolizumab plus Bevacizumab in Unresectable Hepatocellular Carcinoma. *N. Engl. J. Med.* 2020, 382, 1894–1905, doi:10.1056/NEJMoa1915745.
 16. Mokrane, F.-Z.; Lu, L.; Vavasseur, A.; Otal, P.; Peron, J.-M.; Luk, L.; Yang, H.; Ammari, S.; Saenger, Y.; Rousseau, H.; et al. Radiomics Machine-Learning Signature for Diagnosis of Hepatocellular Carcinoma in Cirrhotic Patients with Indeterminate Liver Nodules. *Eur. Radiol.* 2020, 30, 558–570, doi:10.1007/s00330-019-06347-w.
 17. Wu, J.; Liu, A.; Cui, J.; Chen, A.; Song, Q.; Xie, L. Radiomics-Based Classification of Hepatocellular Carcinoma and Hepatic Haemangioma on Precontrast Magnetic Resonance Images. *BMC Med. Imaging* 2019, 19, 23, doi:10.1186/s12880-019-0321-9.
 18. Zhou, Y.; He, L.; Huang, Y.; Chen, S.; Wu, P.; Ye, W.; Liu, Z.; Liang, C. CT-Based Radiomics Signature: A Potential Bi-omarker for Preoperative Prediction of Early Recurrence in Hepatocellular Carcinoma. *Abdom Radiol (NY)* 2017, 42, 1695–1704, doi:10.1007/s00261-017-1072-0.
 19. Wakabayashi, T.; Ouhmich, F.; Gonzalez-Cabrera, C.; Felli, E.; Saviano, A.; Agnus, V.; Savadjiev, P.; Baumert, T.F.; Pessaux, P.; Marescaux, J.; et al. Radiomics in Hepatocellular Carcinoma: A Quantitative Review. *Hepatol. Int.* 2019, 13, 546–559, doi:10.1007/s12072-019-09973-0.
 20. Yip, S.S.F.; Aerts, H.J.W.L. Applications and Limitations of Radiomics. *Phys. Med. Biol.* 2016, 61, R150–66, doi:10.1088/0031-9155/61/13/R150.
 21. Ibrahim, A.; Primakov, S.; Beuque, M.; Woodruff, H.C.; Halilaj, I.; Wu, G.; Refaee, T.; Granzier, R.; Widaatalla, Y.; Hustinx, R.; et al. Radiomics for Precision Medicine: Current Challenges, future Prospects, and the Proposal of a New Framework. *Methods* 2020, doi:10.1016/j.ymeth.2020.05.022.
 22. Larue, R.T.H.M.; van Timmeren, J.E.; de Jong, E.E.C.; Feliciani, G.; Leijenaar, R.T.H.; Schreurs, W.M.J.; Sosef, M.N.; Raat, F.H.P.J.; van der Zande, F.H.R.; Das, M.; et al. Influence of Gray Level Discretization on Radiomic Feature Stability for Different CT Scanners, Tube Currents and Slice Thicknesses: A Comprehensive Phantom Study. *Acta Oncol.* 2017, 56, 1544–1553, doi:10.1080/0284186X.2017.1351624.
 23. van Timmeren, J.E.; Leijenaar, R.T.H.; van Elmpt, W.; Wang, J.; Zhang, Z.; Dekker, A.; Lambin, P. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography* 2016, 2, 361–365, doi:10.18383/j.tom.2016.00208.
 24. Peerlings, J.; Woodruff, H.C.; Winfield, J.M.; Ibrahim, A.; Van Beers, B.E.; Heerschap, A.; Jackson, A.; Wildberger, J.E.; Mottaghy, F.M.; DeSouza, N.M.; et al. Stability of Radiomics Features in Apparent Diffusion Coefficient Maps from a Multi-Centre Test-Retest Trial. *Sci. Rep.* 2019, 9,

- 4800, doi:10.1038/s41598-019-41344-5.
25. Granzier, R.W.Y.; Verbakel, N.M.H.; Ibrahim, A.; van Timmeren, J.E.; van Nijnatten, T.J.A.; Leijenaar, R.T.H.; Lobbes, M.B.I.; Smidt, M.L.; Woodruff, H.C. MRI-Based Radiomics in Breast Cancer: Feature Robustness with Respect to In-ter-Observer Segmentation Variability. *Sci. Rep.* 2020, 10, 14163, doi:10.1038/s41598-020-70940-z.
 26. Leijenaar, R.T.H.; Carvalho, S.; Velazquez, E.R.; van Elmpt, W.J.C.; Parmar, C.; Hoekstra, O.S.; Hoekstra, C.J.; Boellaard, R.; Dekker, A.L.A.J.; Gillies, R.J.; et al. Stability of FDG-PET Radiomics Features: An Integrated Analysis of Test-Retest and Inter-Observer Variability. *Acta Oncol.* 2013, 52, 1391–1397, doi:10.3109/0284186X.2013.812798.
 27. Orlhac, F.; Frouin, F.; Nioche, C.; Ayache, N.; Buvat, I. Validation of a Method to Compensate Multicenter Effects Affecting CT Radiomic Features. 2018.
 28. Orlhac, F.; Boughdad, S.; Philippe, C.; Stalla-Bourdillon, H.; Nioche, C.; Champion, L.; Soussan, M.; Frouin, F.; Frouin, V.; Buvat, I. A Postreconstruction Harmonization Method for Multicenter Radiomic Studies in PET. *J. Nucl. Med.* 2018, 59, 1321–1328, doi:10.2967/jnumed.117.199935.
 29. Da-Ano, R.; Masson, I.; Lucia, F.; Doré, M.; Robin, P.; Alfieri, J.; Rousseau, C.; Mervoyer, A.; Reinhold, C.; Castelli, J.; et al. Performance Comparison of Modified ComBat for Harmonization of Radiomic Features for Multicenter Studies. *Sci. Rep.* 2020, 10, 10248, doi:10.1038/s41598-020-66110-w.
 30. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int. J. Radiat. Oncol. Biol. Phys.* 2018, 102, 1143–1158, doi:10.1016/j.ijrobp.2018.05.053.
 31. Johnson, W.E.; Li, C.; Rabinovic, A. Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods. *Biostatistics* 2007, 8, 118–127, doi:10.1093/biostatistics/kxj037.
 32. Ibrahim, A.; Refaee, T.; Leijenaar, R.T.H.; Primakov, S.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Maidment, A.D.A.; Lambin, P. The Application of a Workflow Integrating the Variable Reproducibility and Harmonizability of Radio-mic Features on a Phantom Dataset. *PLoS One* 2021, 16, e0251147, doi:10.1371/journal.pone.0251147.
 33. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The Effects of in-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* 2021, 13, 1848, doi:10.3390/cancers13081848.
 34. Ricke, J.; Bulla, K.; Kolligs, F.; Peck-Radosavljevic, M.; Reimer, P.; Sangro, B.; Schott, E.; Schütte, K.; Verslype, C.; Wal-ecki, J.; et al. Safety and Toxicity of Radioembolization plus Sorafenib in Advanced Hepatocellular Carcinoma: Analysis of the European Multicentre Trial SORAMIC. *Liver Int.* 2015, 35, 620–626.
 35. World Medical Association World Medical Association Declaration of Helsinki: Ethical Principles for Medical Re-search Involving Human Subjects. *JAMA* 2013, 310, 2191–2194, doi:10.1001/jama.2013.281053.
 36. Zwanenburg, A.; Vallières, M.; Abdalah, M.A.; Aerts, H.J.W.L.; Andrearczyk, V.; Apte, A.; Ashrafinia, S.; Bakas, S.; Beukinga, R.J.; Boellaard, R.; et al. The Image Biomarker Standardization

- Initiative: Standardized Quantitative Radiomics for High-Throughput Image-Based Phenotyping. *Radiology* 2020, 191145, doi:10.1148/radiol.2020191145.
37. Team, R.C. R Language Definition. Vienna, Austria: R foundation for statistical computing 2000.
 38. Gandrud, C. *Reproducible Research with R and R Studio*; CRC Press, 2013; ISBN 9781466572843.
 39. Lin, L.I. A Concordance Correlation Coefficient to Evaluate Reproducibility. *Biometrics* 1989, 45, 255–268.
 40. Stevenson, M.; Stevenson, M.M.; BiasedUrn, I. Package “epiR.” 2020.
 41. Aerts, H.J.W.L.; Velazquez, E.R.; Leijenaar, R.T.H.; Parmar, C.; Grossmann, P.; Carvalho, S.; Bussink, J.; Monshouwer, R.; Haibe-Kains, B.; Rietveld, D.; et al. Decoding Tumour Phenotype by Noninvasive Imaging Using a Quantitative Radiomics Approach. *Nat. Commun.* 2014, 5, 4006, doi:10.1038/ncomms5006.
 42. Perrin, T.; Midya, A.; Yamashita, R.; Chakraborty, J.; Saidon, T.; Jarnagin, W.R.; Gonen, M.; Simpson, A.L.; Do, R.K.G. Short-Term Reproducibility of Radiomic Features in Liver Parenchyma and Liver Malignancies on Contrast-Enhanced CT Imaging. *Abdom Radiol (NY)* 2018, 43, 3271–3278, doi:10.1007/s00261-018-1600-6.
 43. Welch, M.L.; McIntosh, C.; Haibe-Kains, B.; Milosevic, M.F.; Wee, L.; Dekker, A.; Huang, S.H.; Purdie, T.G.; O’Sullivan, B.; Aerts, H.J.W.L.; et al. Vulnerabilities of Radiomic Signature Development: The Need for Safeguards. *Radiother. Oncol.* 2019, 130, 2–9, doi:10.1016/j.radonc.2018.10.027.
 44. Andriarczyk, V.; Depoeringe, A.; Müller, H. Neural Network Training for Cross-Protocol Radiomic Feature Standardization in Computed Tomography. *J Med Imaging (Bellingham)* 2019, 6, 024008, doi:10.1117/1.JMI.6.2.024008.
 45. Bashyam, V.M.; Doshi, J.; Erus, G.; Srinivasan, D.; Abdulkadir, A.; Habes, M.; Fan, Y.; Masters, C.L.; Maruff, P.; Zhuo, C.; et al. Medical Image Harmonization Using Deep Learning Based Canonical Mapping: Toward Robust and Generalizable Learning in Imaging. *arXiv [eess.IV]* 2020.
 46. Modanwal, G.; Vellal, A.; Mazurowski, M.A. Normalization of Breast MRIs Using Cycle-Consistent Generative Adversarial Networks. *arXiv [eess.IV]* 2019.
 47. Dewey, B.E.; Zhao, C.; Reinhold, J.C.; Carass, A.; Fitzgerald, K.C.; Sotirchos, E.S.; Saidha, S.; Oh, J.; Pham, D.L.; Cala-bresi, P.A.; et al. DeepHarmony: A Deep Learning Approach to Contrast Harmonization across Scanner Changes. *Magn. Reson. Imaging* 2019, 64, 160–170, doi:10.1016/j.mri.2019.05.041.

A large, stylized white number 9 is centered on a blue watercolor splash. The splash is composed of various shades of blue, from light to dark, with some darker spots and a textured, organic appearance. The number 9 is a simple, clean, sans-serif font. The background is white.

9

Chapter 9

MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability

Authors

R. Granzier, N.M.H. Verbakel, A. Ibrahim, J.E. van Timmeren, T.J.A. van Nijnatten,
R.T.H. Leijenaar, M.B.I. Lobbes, M.L. Smidt and H.C. Woodruff

Adapted from

Scientific reports. 2020 Aug 25;10(1):1-1

DOI

10.1038/s41598-020-70940-z

Abstract

Radiomics is an emerging field using the extraction of quantitative features from medical images for tissue characterization. While MRI-based radiomics is still at an early stage, it showed some promising results in studies focusing on breast cancer patients in improving diagnoses and therapy response assessment. Nevertheless, the use of radiomics raises a number of issues regarding feature quantification and robustness. Therefore, our study aim was to determine the robustness of radiomics features extracted by two commonly used radiomics software with respect to variability in manual breast tumor segmentation on MRI. A total of 129 histologically confirmed breast tumors were segmented manually in three dimensions on the first post-contrast T1-weighted MR exam by four observers: a dedicated breast radiologist, a resident, a Ph.D. candidate, and a medical student. Robust features were assessed using the intraclass correlation coefficient (ICC >0.9). The inter-observer variability was evaluated by the volumetric Dice Similarity Coefficient (DSC). The mean DSC for all tumors was 0.81 (range 0.19-0.96), indicating a good spatial overlap of the segmentations based on observers of varying expertise. In total, 41.6% (552/1328) and 32.8% (273/833) of all RadiomiX and Pyradiomics features, respectively, were identified as robust and were independent of inter-observer manual segmentation variability.

Introduction

Radiomics is a technique that is used to extract large amounts of quantitative information from routine medical images that decode information about a region of interest (ROI). The majority of radiomics articles published concerns its application in the oncological field⁽¹⁻⁴⁾. Here, radiomics bears the advantage of non-invasively quantifying the underlying phenotype of the entire tumor for multiple lesions simultaneously, in contrast to tissue biopsy, which samples only a small part of a single (often heterogeneous) tumor^(2, 5). The ability to characterize the tumor and to establish links to the underlying biology⁽⁶⁾ and ultimately clinical outcomes, allows a more patient-tailored treatment⁽⁷⁾, enabling ‘precision medicine’^(8, 9). Recently, several articles have outlined the potential clinical applicability of radiomics in the field of breast cancer for different purposes, *e.g.* diagnosis^(10, 11), tumor response prediction⁽¹²⁻¹⁴⁾, prediction of molecular tumor subtype^(15, 16), and prediction of axillary lymph node metastases^(17, 18).

Although these results are promising, issues regarding features robustness as well as the comparability of results, including inter-observer segmentation variability, need to be addressed⁽¹⁹⁻²⁴⁾. In order to extract clinically useful information from medical images and to use features as clinical biomarkers, it is important that extracted features are reproducible, standardized and robust^(25, 26). All consecutive steps in the radiomics workflow induce potential uncertainties regarding feature robustness^(27, 28). Since there used to be no gold standard or guideline for extraction of image features for radiomics use, an initiative –Image Biomarker Standardization Initiative (IBSI)- was launched as an effort to standardize the entire radiomics extraction process and encourage feature robustness⁽²⁹⁾.

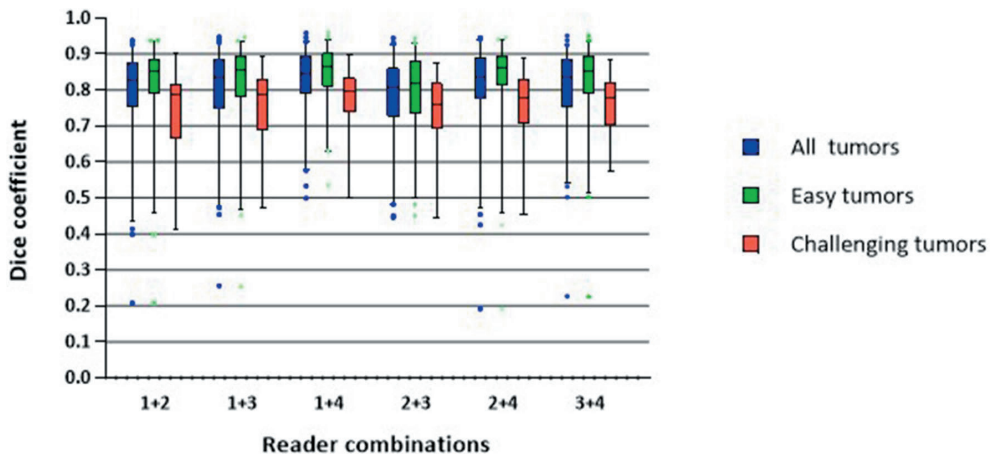


Figure I: Tumor segmentation variability for pairwise comparison of the different observers. 1) Dedicated breast radiologist, 2) Radiology resident, 3) Ph.D. candidate with a medical degree and 4) Medical student.

ROI segmentation is an important step after image acquisition in the radiomics workflow, and one of the largest bottlenecks⁽³⁰⁾. Traditionally, the edges (2D) or surfaces (3D) of the ROI are segmented, thereby defining a region from which features will be extracted. Segmentation can be performed either manually, semi-automatically, or completely automatically. Both manual and semi-automatic segmentation are prone to inter- and intra-observer variabilities, with the degree of observer experience playing an important role⁽³¹⁻³³⁾.

To the best of our knowledge, no articles have been published on the effect of manual inter-observer segmentation variability on MRI-based feature robustness in breast cancer patients. MRI is the most accurate modality for neoadjuvant systemic therapy response monitoring in breast cancer patients and as such much used in daily clinical practice⁽³⁴⁻³⁷⁾. In this article, we investigate the robustness of MR radiomics features, extracted using two commonly used radiomics software, with respect to variations in manual tumor segmentation of breast cancer patients.

Results

Study population

After the application of inclusion and exclusion criteria, 102 patients were included in the final analysis. Twenty-one of these patients were diagnosed with multifocal breast cancer, bringing the total number of tumors analyzed in this study to 129. Of these, 94 tumors (73%) were assigned 'easy tumors' and the remaining 35 tumors (27%) were assigned 'challenging tumors'. The tumor volume between both groups was significant differently (5.3 vs 10.4 for 'easy and challenging tumors', respectively, $p=0.03$)

Segmentation variability

DSC distributions of all observer combinations are shown in Figure 1. The mean DSC was 0.81 (range 0.19-0.96). The mean DSC was higher for the 'easy tumors' compared to the 'challenging tumors' (0.83 vs. 0.75, respectively, $p<0.001$). The mean DSC for each observer combination separately, for all tumors, ranged between 0.78 and 0.83, where the segmentations of the breast radiologist and the medical student showed the highest overlap.

Pre-processing and feature extraction

The bin width for image discretization (calculated from the ROI greyscale range) was 0.1. Discretization of the scans with bins 0.1 wide resulted in a mean of 61 grayscale values per image (range 27 -131). RadiomiX and Pyradiomics software extracted a total of 1328 and 833 features for each ROI, respectively. The extracted radiomics features included shape features, first-order statistical, intensity-histogram based, fractal, local

intensity, and texture matrix-based features from both unfiltered and filtered images (wavelet decompositions). The RadiomiX software extracts more feature groups compared to the Pyradiomics software, namely intensity histogram (IH), fractal, local intensity, and gray level dependency zone matrix (GLDZM) features.

Table 1: average ICC values per feature group of the unfiltered and wavelet RadiomiX and Pyradiomics features .

Feature group (n)	OncoRadiomiX		Pyradiomics	
	Mean ICC	Range	Mean ICC	Range
Shape	0.79	0.57 – 0.93	0.80	0.69 – 0.92
Signal intensity				
- First-order statistics	0.85	0.51 – 0.99	0.84	0.50 – 0.97
- IH	0.76	0.63 – 0.98	-	-
Fractal	0.81	0.79 – 0.83	-	-
LocInt	0.95	0.93 – 0.96	-	-
GLCM	0.76	0.49 – 0.88	0.80	0.71 – 0.88
GLRLM	0.79	0.56 – 0.96	0.81	0.63 – 0.95
GLSZM	0.80	0.55 – 0.98	0.84	0.58 – 0.97
GLDZM	0.76	0.50 – 0.92	-	-
NGTDM	0.78	0.57 – 0.85	0.80	0.72 – 0.91
(N)GLDM	0.83	0.55 – 0.96	0.79	0.52 – 0.96
Wavelet	0.81	0.01 – 0.99	0.81	0.12 – 0.99

Radiomics feature robustness

The average ICC for all RadiomiX features was 0.86 (95% CI: 0.85-0.86) and for all Pyradiomics features 0.84 (95% CI: 0.83-0.84). Table 1 presents the average ICC value per feature group for both software. The local intensity features scored the highest average ICC value for the RadiomiX features, and the first-order statistical features score the highest average ICC for the Pyradiomics features.

The percentage of features that scored an ICC > 0.90, and thus were labeled by our pre-determined ICC cut-off as robust, was 41.6% (552/1328) for RadiomiX features and 32.8% (273/833) for Pyradiomics features. The unfiltered RadiomiX features (*i.e.*, calculated on the unfiltered images) had an average ICC value of 0.79 (95% CI: 0.77 – 0.81), of which 41.1% (69/168) were robust (Figure 2). The unfiltered Pyradiomics features had an average ICC value of 0.81 (95% CI: 0.79-0.83), of which 16.2% (17/105) were robust (Figure 3). The results of the wavelet feature groups for both software are presented in the supplementary material 1 and 2.

The percentage of robust RadiomiX features for the ‘easy tumors’ and the ‘challenging

tumors' was 57.5% (763/1328) and 17.2% (228/1328), respectively. When only considering the 168 unfiltered features, 50.0% (84/168) of the 'easy tumors' were robust and 20.2% (34/168) of the 'challenging tumors' (supplementary material 3). The percentage of robust Pyradiomics features for the 'easy tumors' and the 'challenging tumors' was 35.7% (297/833) and 28.6% (238/833), respectively. When only considering the 105 unfiltered features, 23.8% (25/105) of the 'easy tumors' were robust and 14.3% (15/105) of the 'challenging tumors' (supplementary material 4).

Discussion

In this study, our ultimate goal was to define a list of robust MRI radiomics features, independent of inter-observer segmentation variability, which could facilitate further breast MRI-based radiomics research. We successfully identified a subgroup of robust features for two commonly used radiomics software (41.6% of all RadiomiX features and 32.8% of all Pyradiomics features) in the presence of inter-observer segmentation variability (mean DSC of 0.81).

Although MRI feature robustness has already been investigated for different tumor sites (*e.g.*, cervical cancer ⁽¹⁹⁾ and glioblastoma ⁽²³⁾), the effect of inter-observer variability segmentation is most likely tumor-site specific ⁽³⁸⁾. The feature groups enclosing the most robust features in previous investigations (shape ⁽¹⁹⁾ and, Intensity-histogram and GLCM ⁽²³⁾) are different from what we found to be the feature group enclosing the most robust features (local intensities and GLRLM). Most likely this could be explained that different tumor sites influence inter-observer variability. Although one must not forget that the differences in MRI sequences and, feature extraction software also influence this variability. Therefore, the MRI feature robustness cannot be generalized and must be examined for each specific tumor site, taking into account different MRI sequences and feature extraction software.

In addition, feature robustness for both radiomics software was identified for 'easy tumors' and 'challenging tumors'. The number of robust features increased for 'easy tumors' and decreased for 'challenging tumors' in both software with significant differences between the mean DSC of the 'easy' and 'challenging' tumors (0.83 vs. 0.75, respectively, $p < 0.001$). The fact that the 'challenging tumors' were more irregular, often with spiculae, causes more segmentation variability and therefore less robust features. Furthermore, the significant difference in the DSC between easy and challenging tumors could be attributed to the sensitivity of the metric to tumor volume. Easy tumors were on average significantly smaller than challenging ones; therefore, a minor difference in segmentation of a small tumor would have a more profound effect on the DSC, compared to those with larger volumes.



Figure 2: ICC values of all unfiltered RadiomiX features with robust features (ICC > 0.90) shown in green.

Chapter 9



Figure 3: ICC values of all unfiltered Pyradiomics features with robust features (ICC > 0.90) shown in green.

A detailed comparison to previous studies is limited to one similar study. Saha et al⁽³⁹⁾ investigated the impact of breast MRI segmentation variability on radiomics feature robustness, whereby features were extracted using in-house software. Their reported mean ICC of 0.85 for all features, using semi-automatic breast tumor segmentation, is comparable to the average ICC reported in this study. Although the segmentations were performed by four fellow breast radiology trainees, the DSC results they report (range: 0.506-0.740) were much lower than the DSC results in our analysis (range: 0.783-0.827). We consciously opted for people with different segmentation expertise to ensure observer-independence of the robust features, consequently widening the applicability. Approximately 10% of the tumor features in their article were found to be robust, compared to 41.1% in this study. Solely 20 textural features (GLCM) were comparable between the studies, whereby the ICC of these features showed a substantial difference (average 0.26, range 0.09 – 0.51).

While we present the robust features for two different radiomics software, our aim is solely to facilitate future application of our findings. Both software have different pre-processing steps, and different groups of features, and comparing the software is beyond the scope of this study. A global initiative to standardize radiomic features extraction using different radiomics software—Imaging Biomarkers Standardization Initiative (IBSI)- was started to address these issues in a more comprehensive fashion⁽⁴⁰⁾.

To overcome the problem of inter-observer variability with respect to ROI segmentation, promising steps towards (semi-)automatic segmentation have been taken in other tumor sites⁽⁴¹⁻⁴⁵⁾. However, little work has been published on fully automatic segmentation software for DCE-MRI of the breast^(33, 46-48). Most software, including semi-automatic segmentation, still require manual input or adjustments^(33, 46, 47), and would still be significantly slower than fully automated segmentation. Recent work on automatic MRI breast tissue segmentation reported encouraging results but was performed on only 30 patients⁽⁴⁸⁾. The current lack of reliable, validated and widely available automatic segmentation software tools, and the need for manual input in semi-automated segmentation, demonstrate that manual segmentation remains important. The use of protocols or guidelines could encourage more reproducible manual segmentation results^(49, 50). Furthermore, by providing precise instructions before the start of segmentation, inter-observer segmentation variability can be minimized.

There are some limitations to this study. Although an ICC threshold value of 0.90 was chosen to determine feature robustness, the significance of this threshold for radiomics models for patients' outcome prediction is yet to be investigated. The inclusion of more patients and observers will allow better generalization of the results and development of robust radiomics signatures. Furthermore, we identified feature robustness to

segmentation observer variability. However, due to the lack of data, we were not able to assess the robustness of radiomics features to differences in image acquisition, pre-processing and feature extraction, which are other major challenges in radiomics analysis. These are the aim of our current studies.

In conclusion, this study shows the intuitive notion that more complex, challenging tumors lead to less robust features. We identified radiomics features robust to inter-observer variations across two different radiomics software, which could be used for preselection of radiomics features in future radiomics analysis concerning MRI-based breast radiomics. Ultimately, this study identified a list of robust radiomics features, which is independent of inter-observer segmentation variability in breast MRI for two commonly used software.

Material and Methods

Study population

In this single-center retrospective study, we collected data on 138 patients with histologically confirmed invasive breast cancer, who were planned for receiving NST and underwent a pretreatment DCE-MRI between January 2011 and December 2017 in Maastricht University Medical Center+. The institutional research board of the MUMC+ approved the study and waived the requirement for informed consent and the further need of guidelines. Exclusion criteria were: pathologically confirmed mastitis carcinomatosa, MR scan artifacts, or refusal of medical record usage by the patient. Furthermore, we excluded patients that underwent breast MRI exams with non-standard acquisition parameters, due to the use of a different MR scanner. All histologically confirmed breast tumors were included in the analysis. The complete process is summarized in the flowchart presented in Figure 4.

Imaging data

All images were acquired by two clinically interchangeable (i.e. provide qualitatively similar images) 1.5T MRI scanners (Philips Intera and Philips Ingenia), using a dynamic contrast-enhanced T1-weighted (DCE-T1W) sequence with similar acquisition protocols (Table 2). The patients were scanned in prone position using a 16-channel dedicated breast coil. The DCE-T1W images were obtained before and after intravenous injection of gadolinium-based contrast Gadobutrol (Gadovist (EU)) with a volume of 15 cc and a flow rate of 2 ml/sec. One pre-contrast image and five post-contrast images were obtained for each patient.

Tumor segmentation

The T1W images acquired two minutes post-contrast administration were used for the 3D tumor segmentation, as this is generally accepted to be the peak of enhancement of breast cancers (51). Tumors were independently segmented by four observers with different degrees of experience in breast MR imaging: a dedicated breast radiologist with 11 years of clinical breast MRI experience (ML), a radiology resident with one year of breast MRI clinical experience (TvN), a Ph.D. candidate with a medical degree but no breast MRI clinical experience (RG) and a medical student with no experience whatsoever (NV) (Figure 5). Segmentations were performed manually with Mirada RTx (v1.2.0.59, Mirada Medical, Oxford, UK). Agreements regarding segmentation procedures were made prior to tumor segmentation: (i) observers were allowed to adjust the image grayscale to optimize the visualization of the tumor; (ii) lymph nodes, pectoral muscle, and skin were excluded from segmentation; (iii) spiculae were only segmented if histologically confirmed. All observers had access to the radiology report during segmentation but were blinded to each other's segmentations.

Image pre-processing and feature extraction

Radiomics feature extraction is generally performed after image pre-processing. Pre-processing is designed to increase data homogeneity, as well as to reduce image noise and computational requirements. Both radiomics software have the optionality to perform image normalization internally before feature extraction, which varies to an extent across the software. Pyradiomics centers the image around the mean and standard deviation based on all gray values of the image, while RadiomiX normalizes the images after removal of background data (non-breast voxels containing air). This transforms the voxel grayscale values to a more comparable range without changing image textures. Each image was discretized by resampling the grayscale values using a fixed bin width of 0.1 in order to reduce image noise and computational burden. The Pyradiomics community⁽⁵²⁾ recommends the number of bins to be in range of 16-128. We calculated the optimal bin width by extracting the greyscale ranges within all the ROIs and choosing a width that maximizes the number of ROIs that fall in the abovementioned range of bins. Finally, voxel size was standardized across the cohorts to isotropic 1.0 mm³ voxels by means of linear interpolation. For each manually segmented ROI, features were extracted using two commonly used radiomics software: RadiomiX Discovery Toolbox software (OncoRadiomics SA, Liège, Belgium) and the open-source Pyradiomics software, version 2.1.2^(52,53). A mathematical description of all RadiomiX features can be found in supplementary material 5. The Pyradiomics feature description can be found online⁽⁵⁴⁾. Both software are IBSI compliant for most features, with a note being added in case of differences.

Table 2: Imaging parameters for the breast DCE TiW sequence for both scanners.

	Scanner 1 Philips Ingenia (n)	Scanner 2 Philips Intera (n)
Number of tumors	100	29
Field strength (T)	1.5	1.5
Slice thickness (mm)	1.0	1.0
Repetition time (msec)	7.5 (88), 7.6 (12)	7.4 (13), 7.5 (15), 7.6 (1)
Echo time (msec)	3.4	3.4
Flip angle (degrees)	10	10
Echo train length	89* (range 62-175)	80* (range 60 – 85)
Pixel spacing (mm)	0.79 ² (3), 0.85 ² (1), 0.92 ² (2), 0.95 ² (47), 0.95 ² (47)	0.85 ² (1), 0.94 ² (1), 0.97 ² (26), 0.99 ² (1)
Temporal resolution (sec)	95	98

*average.

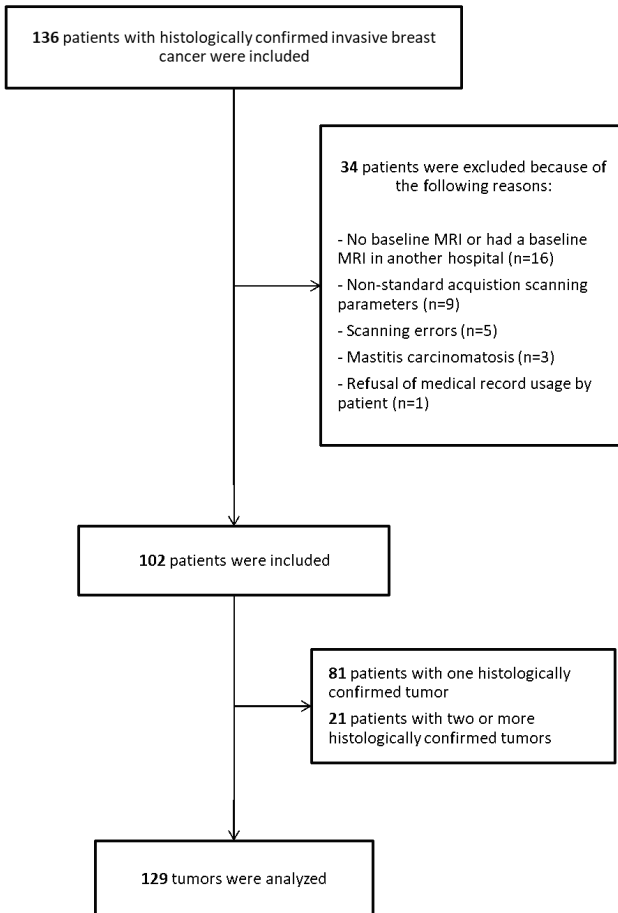


Figure 4: Flowchart of the patient population in the study .

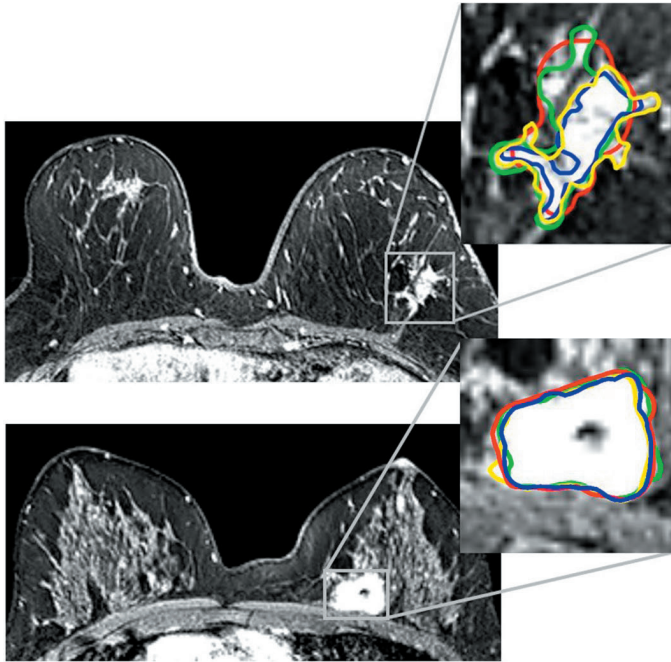


Figure 5: Two invasive breast tumors in the left breast on the 2-minute post-contrast DCE-MRI with four single manual segmentations (colored margins: red, blue, green and yellow) fused. Upper: ‘challenging tumor’ with a mean DSC of 0.78 (range 0.71 - 0.82). Lower: ‘easy tumor’ with a mean DSC of 0.90 (range 0.89 - 0.91).

Data analysis

Segmentation variability analysis

Features with (near) zero variance across all tumors, *i.e.* features that have the same value across ninety-five percent or more of the observations, were excluded from the analysis as they carry no discriminative value. To evaluate the variability of the remaining features introduced by manual segmentation, the volumetric Dice Similarity Coefficient (DSC) was calculated for all pairs of observers. The DSC is a metric that quantifies the agreement (or ‘overlap’) between two segmentations⁽⁵⁵⁾. A DSC of 1 indicates perfect spatial overlap of the segmentations, whereas 0 indicates no agreement, *i.e.* no spatial overlap of the segmentations, and a good overlap is considered with $DSC > 0.7$ as indicated by the literature⁽⁵⁶⁾. The DSC was calculated as:

$$DSC = 2 \frac{|A \cap B|}{(|A| + |B|)}$$

where A is the set of voxels contained in the first contour, B is the set of voxels contained in the second contour, $||$ indicates the cardinality of the sets, and \cap is the intersection between the first and second sets⁽⁵⁷⁾. The DSC was calculated using Python (Version 3.6.3150.1013).

Radiomics feature robustness analysis

Feature robustness was assessed by evaluating the two-way random single measure intraclass correlation coefficient (ICC) ^(2,1). The two-way random model approach was chosen as it allows generalization of the results to any other rater with similar characteristics ⁽⁵⁷⁾. The ICC ranges between 0 and 1, with values closer to 1 representing stronger feature robustness to differences in segmentations. We chose a pre-defined ICC cut-off of >0.9 to select highly stable features that are insensitive to segmentation variability ⁽⁵⁷⁾. Feature robustness was calculated for all RadiomiX and Pyradiomics features. The settings for image pre-processing (normalization, discretization, and resampling) in both radiomics software were checked for disparities. Calculations were performed in R studio (version 1.1.456, Vienna, Austria) ⁽⁵⁸⁾ using the IRR package version 0.84 ⁽⁵⁹⁾.

Easy- vs. challenging-to-segment tumors analysis

The differences in feature robustness and inter-observer tumor segmentation variability between ‘easy-to-segment’ and ‘challenging-to-segment’ tumors ones, hereinafter referred to as ‘easy tumors’ and ‘challenging tumors’, were assessed. This classification was unanimously determined by the dedicated breast radiologist (ML). ‘Easy tumors’ were defined as homogenous, round tumors with relatively sharp (albeit sometimes irregular) margins, without spiculae or areas of accompanying non-mass enhancement. Tumors not meeting these criteria were categorized as ‘challenging tumors’ (Figure 5). To compare DSC results between ‘easy’ and ‘challenging’ tumors we used the independent samples t-test, performed in R studio using the IRR package.

Acknowledgements

This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018-2

Authors’ contribution

Hereby an original research manuscript is submitted to Scientific Reports journal. All authors have read the present manuscript and approved to submit. All authors have made a substantial contribution to the manuscript. R.G. and N.V., A.I wrote the manuscript, which was checked and revised several times by all authors. R.G., M.S., and T.v.N. contributed to the concept of the manuscript. Specific work on the design of the manuscript and figure preparation was performed by R.G., A.I., J.v.T., R.L. and H.W. Tumor segmentations were performed by R.G., M.L., T.v.N., and N.V. Data analysis and interpretation was performed by R.G., N.V., A.I and H.W.

Competing interests

Dr. Woodruff and Dr. Leijenaar have (minority) shares in the company OncoRadiomics. Dr. Smidt received a grant of the company Servier for microbiome research. The rest of authors declare no competing interest.

References

1. Lambin P, Rios-Velazquez E, Leijenaar R, Carvalho S, van Stiphout RG, Granton P, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *European journal of cancer (Oxford, England : 1990)*. 2012;48(4):441-6.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
3. Gillies R, Kinahan P, Hricak H. Radiomics: Images are more than pictures, they are data. *Radiology*. 2016;278(2):563-77.
4. Bogowicz M, Vuong D, Huellner MW, Pavic M, Andratschke N, Gabrys HS, et al. CT radiomics and PET radiomics: ready for clinical implementation? *Q J Nucl Med Mol Imaging*. 2019.
5. Davnall F, Yip CS, Ljungqvist G, Selmi M, Ng F, Sanghera B, et al. Assessment of tumor heterogeneity: an emerging imaging tool for clinical practice? *Insights Imaging*. 2012;3(6):573-89.
6. Grossmann P, Stringfield O, El-Hachem N, Bui MM, Rios Velazquez E, Parmar C, et al. Defining the biological basis of radiomic phenotypes in lung cancer. *Elife*. 2017;6.
7. Ibrahim A, Vallières M, Woodruff H, Primakov S, Beheshti M, Keek S, et al., editors. *Radiomics Analysis for Clinical Decision Support in Nuclear Medicine*. *Seminars in Nuclear Medicine*; 2019: Elsevier.
8. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-62.
9. Walsh S, de Jong EE, van Timmeren JE, Ibrahim A, Compter I, Peerlings J, et al. Decision support systems in oncology. *JCO clinical cancer informatics*. 2019;3:1-9.
10. Milenkovic J, Dalmis MU, Zgajnar J, Platel B. Textural analysis of early-phase spatiotemporal changes in contrast enhancement of breast lesions imaged with an ultrafast DCE-MRI protocol. *Med Phys*. 2017;44(9):4652-64.
11. Parekh VS, Jacobs MA. Integrated radiomic framework for breast cancer and tumor biology using advanced machine learning and multiparametric MRI. *NPJ Breast Cancer*. 2017;3:43.
12. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res*. 2019.
13. Xiong Q, Zhou X, Liu Z, Lei C, Yang C, Yang M, et al. Multiparametric MRI-based radiomics analysis for prediction of breast cancers insensitive to neoadjuvant chemotherapy. *Clin Transl Oncol*. 2019.
14. Cain EH, Saha A, Harowicz MR, Marks JR, Marcom PK, Mazurowski MA. Multivariate machine learning models for prediction of pathologic response to neoadjuvant therapy in breast cancer using MRI features: a study using an independent validation set. *Breast cancer research and treatment*. 2018.
15. Waugh SA, Purdie CA, Jordan LB, Vinnicombe S, Lerski RA, Martin P, et al. Magnetic resonance

- imaging texture analysis classification of primary breast cancer. *Eur Radiol.* 2016;26(2):322-30.
16. Monti S, Aiello M, Incoronato M, Grimaldi AM, Moscarino M, Mirabelli P, et al. DCE-MRI Pharmacokinetic-Based Phenotyping of Invasive Ductal Carcinoma: A Radiomic Study for Prediction of Histological Outcomes. *Contrast Media Mol Imaging.* 2018;2018:5076269.
 17. Cui X, Wang N, Zhao Y, Chen S, Li S, Xu M, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer using Radiomics Features of DCE-MRI. *Sci Rep.* 2019;9(1):2240.
 18. Yang J, Wang T, Yang L, Wang Y, Li H, Zhou X, et al. Preoperative Prediction of Axillary Lymph Node Metastasis in Breast Cancer Using Mammography-Based Radiomics Method. *Sci Rep.* 2019;9(1):4429.
 19. Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology.* 2019;135:107-14.
 20. Qiu Q, Duan J, Duan Z, Meng X, Ma C, Zhu J, et al. Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability. *Quant Imaging Med Surg.* 2019;9(3):453-64.
 21. Belli ML, Mori M, Broggi S, Cattaneo GM, Bettinardi V, Dell'Oca I, et al. Quantifying the robustness of [(18)F]FDG-PET/CT radiomic features with respect to tumor delineation in head and neck and pancreatic cancer patients. *Phys Med.* 2018;49:105-11.
 22. Pavic M, Bogowicz M, Wurms X, Glatz S, Finazzi T, Riesterer O, et al. Influence of inter-observer delineation variability on radiomics stability in different tumor sites. *Acta Oncol.* 2018;57(8):1070-4.
 23. Tixier F, Um H, Young RJ, Veeraraghavan H. Reliability of tumor segmentation in glioblastoma: Impact on the robustness of MRI-radiomic features. *Med Phys.* 2019.
 24. Traverso A, Kazmierski M, Shi Z, Kalendralis P, Welch M, Nissen HD, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Phys Med.* 2019;61:44-51.
 25. Rizzo S, Botta F, Raimondi S, Origgi D, Fanciullo C, Morganti AG, et al. Radiomics: the facts and the challenges of image analysis. *Eur Radiol Exp.* 2018;2(1):36.
 26. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and Reproducibility of Radiomic Features: A Systematic Review. *Int J Radiat Oncol Biol Phys.* 2018;102(4):1143-58.
 27. Larue RT, Defraene G, De Ruyscher D, Lambin P, van Elmpt W. Quantitative radiomics studies for tissue characterization: a review of technology and methodological procedures. *Br J Radiol.* 2017;90(1070):20160665.
 28. Yip SSF, Aerts HJWL. Applications and limitations of radiomics. *Phys Med Biol.* 2016;61(13):R150-R66.
 29. Zwanenburg A LS, Vallières M, Lock S. Image biomarker standardisation initiative. . arXiv preprint arXiv:161207003.
 30. Polan DF, Brady SL, Kaufman RA. Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study. *Phys Med Biol.* 2016;61(17):6553-69.
 31. Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *J Med Phys.* 2008;33(4):136-40.

32. Beresford MJ, Padhani AR, Taylor NJ, Ah-See ML, Stirling JJ, Makris A, et al. Inter- and intraobserver variability in the evaluation of dynamic breast cancer MRI. *J Magn Reson Imaging*. 2006;24(6):1316-25.
33. Saha A, Grimm LJ, Harowicz M, Ghate SV, Kim C, Walsh R, et al. Interobserver variability in identification of breast tumors in MRI and its implications for prognostic biomarkers and radiogenomics. *Med Phys*. 2016;43(8):4558.
34. Lobbes MB, Prevos R, Smidt M, Tjan-Heijnen VC, van Goethem M, Schipper R, et al. The role of magnetic resonance imaging in assessing residual disease and pathologic complete response in breast cancer patients receiving neoadjuvant chemotherapy: a systematic review. *Insights Imaging*. 2013;4(2):163-75.
35. Houssami N, Turner R, Morrow M. Preoperative magnetic resonance imaging in breast cancer: meta-analysis of surgical outcomes. *Annals of surgery*. 2013;257(2):249-55.
36. Woolf DK, Padhani AR, Taylor NJ, Gogbashian A, Li SP, Beresford MJ, et al. Assessing response in breast cancer with dynamic contrast-enhanced magnetic resonance imaging: are signal intensity-time curves adequate? *Breast Cancer Res Treat*. 2014;147(2):335-43.
37. Cardoso F, Kyriakides S, Ohno S, Penault-Llorca F, Poortmans P, Rubio IT, et al. Early breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*. 2019.
38. van Timmeren JE, Leijenaar RTH, van Elmpt W, Wang J, Zhang Z, Dekker A, et al. Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific? *Tomography*. 2016;2(4):361-5.
39. Saha A, Harowicz MR, Mazurowski MA. Breast cancer MRI radiomics: An overview of algorithmic features and impact of inter-reader variability in annotating tumors. *Med Phys*. 2018;45(7):3076-85.
40. Zwanenburg A, Leger S, Vallières M, Löck S. Image biomarker standardisation initiative. *arXiv preprint arXiv:161207003*. 2016.
41. Hong J, Park BY, Lee MJ, Chung CS, Cha J, Park H. Two-step deep neural network for segmentation of deep white matter hyperintensities in migraineurs. *Comput Methods Programs Biomed*. 2019;183:105065.
42. Ghavami N, Hu Y, Gibson E, Bonmati E, Emberton M, Moore CM, et al. Automatic segmentation of prostate MRI using convolutional neural networks: Investigating the impact of network architecture on the accuracy of volume measurement and MRI-ultrasound registration. *Med Image Anal*. 2019;58:101558.
43. Kugelman J, Alonso-Caneiro D, Read SA, Hamwood J, Vincent SJ, Chen FK, et al. Automatic choroidal segmentation in OCT images using supervised deep learning methods. *Sci Rep*. 2019;9(1):13298.
44. Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, et al. Rapid-learning system for cancer care. *Journal of Clinical Oncology*. 2010;28(27):4268.
45. Heye T, Merkle EM, Reiner CS, Davenport MS, Horvath JJ, Feuerlein S, et al. Reproducibility of dynamic contrast-enhanced MR imaging. Part II. Comparison of intra- and interobserver variability with manual region of interest placement versus semiautomatic lesion segmentation and histogram

- analysis. *Radiology*. 2013;266(3):812-21.
46. Chen W, Giger ML, Bick U. A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Acad Radiol*. 2006;13(1):63-72.
 47. Lin MQ, Chen JH, Wang XY, Chan SW, Chen SP, Su MY. Template-based automatic breast segmentation on MRI by excluding the chest region. *Med Phys Journal Translated Name Medical Physics*. 2013;40(12).
 48. Thakran S, Chatterjee S, Singhal M, Gupta RK, Singh A. Automatic outer and inner breast tissue segmentation using multi-parametric MRI images of breast tumor patients. *PLoS One*. 2018;13(1):e0190348.
 49. Fuller CD, Nijkamp J, Duppen JC, Rasch CR, Thomas CR, Jr., Wang SJ, et al. Prospective randomized double-blind pilot study of site-specific consensus atlas implementation for rectal cancer target volume delineation in the cooperative group setting. *Int J Radiat Oncol Biol Phys*. 2011;79(2):481-9.
 50. Mitchell DM, Perry L, Smith S, Elliott T, Wylie JP, Cowan RA, et al. Assessing the effect of a contouring protocol on postprostatectomy radiotherapy clinical target volumes and interphysician variation. *Int J Radiat Oncol Biol Phys*. 2009;75(4):990-3.
 51. El Khouli RH, Macura KJ, Jacobs MA, Khalil TH, Kamel IR, Dwyer A, et al. Dynamic contrast-enhanced MRI of the breast: quantitative method for kinetic curve type assessment. *AJR Am J Roentgenol*. 2009;193(4):W295-300.
 52. Pyradiomics feature description [Available from: <https://pyradiomics.readthedocs.io/en/latest/features.html>].
 53. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e7.
 54. Ramesh A, Kambhampati C, Monson JR, Drew P. Artificial intelligence in medicine. *Annals of The Royal College of Surgeons of England*. 2004;86(5):334.
 55. Measures of the Amount of Ecologic Association Between Species. *Ecology*. 1945;26(3):297-302.
 56. Zijdenbos AP, Dawant BM, Margolin RA, Palmer AC. Morphometric analysis of white matter lesions in MR images: method and validation. *IEEE transactions on medical imaging*. 1994;13(4):716-24.
 57. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J Chiropr Med*. 2016;15(2):155-63.
 58. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics*. 2012;27(1):167-72.
 59. Gamer M, Lemon J, Fellows I, Sing P. Various Coefficients of Interrater Reliability and Agreement. *IRR: R package version 084*. 2012.



10

Chapter 10

Test-retest data for the assessment of breast MRI radiomic feature repeatability

Authors

R.W.Y. Granzier, A. Ibrahim, S. Primakov, S.A. Keek, I. Halilaj, A. Zwanenburg, S.M.E. Engelen, M.B.I. Lobbes, P. Lambin, H.C. Woodruff and M.L. Smidt

Adapted from

Journal of Magnetic Resonance Imaging. 2021 Dec 22

DOI

10.1002/jmri.28027

Abstract

Background: Radiomic features extracted from breast MRI have potential for diagnostic, prognostic, and predictive purposes. However, before they can be used as biomarkers in clinical decision support systems, features need to be repeatable and reproducible.

Objective: Identify repeatable radiomics features within breast tissue on prospectively collected MRI exams through multiple test-retest measurements.

Study type: Prospective

Population: 11 healthy female volunteers

Field strength/sequence: 1.5 T;MRI exams, comprising T2-weighted turbo spin-echo (T2W) sequence, native T1-weighted turbo gradient-echo (T1W) sequence, diffusion-weighted imaging (DWI) sequence using b-values 0/150/800, and corresponding derived ADC maps.

Assessment: 18 MRI exams (three test-retest settings, repeated on two days) per healthy volunteer were examined on an identical scanner using a fixed clinical breast protocol. For each scan, 91 features were extracted from the 3D manually segmented right breast using Pyradiomics, before and after image pre-processing. Image pre-processing consisted of (i) bias field correction (BFC), (ii) z-score normalization with and without BFC, (iii) grayscale discretization using 32 and 64 bins with and without BFC, and (iv) z-score normalization + grayscale discretization using 32 and 64 bins with and without BFC.

Statistical tests: Features' repeatability was assessed using concordance correlation coefficient(CCC) for each pair, i.e. each MRI was compared to each of the remaining 17 MRI with a cut-off value of $CCC > 0.90$.

Results: Images without pre-processing produced the highest number of repeatable features for both T1W sequence and ADC maps with 15/91 (16.5%) and 8/91 (8.8%) repeatable features, respectively. Pre-processed images produced between 4/91 (4.4%) and 14/91 (15.4%), and 6/91 (6.6%) and 7/91 (7.7%) repeatable features, respectively for T1W and ADC maps. Z-score normalization produced highest number of repeatable features, 26/91 (28.6%) in T2W sequences, in these images, no pre-processing produced 11/91 (12.1%) repeatable features.

Data conclusion: Radiomic features extracted from T1W, T2W sequences and ADC maps from breast MRI exams showed a varying number of repeatable features, depending on the sequence. Effects of different preprocessing procedures on repeatability of features were different for each sequence.

Keywords: Breast, MRI, Radiomics, Feature repeatability

Introduction

The use of radiomics to answer diagnostic, predictive, and prognostic questions has increased in recent years, especially in the field of oncology⁽¹⁾. Radiomics refers to the extraction of large amounts of high-throughput quantitative data from medical images using mathematical algorithms that have the potential to noninvasively reveal more information about the region of interest than can be captured by visual inspection alone⁽²⁾. The extracted quantitative data, termed radiomics features, capture information regarding the shape, intensity, and texture of the chosen region of interest (ROI), which is usually the lesion or the affected organ. Radiomics features are intended to serve as biomarkers for the development of clinical decision support systems to enhance personalized medicine⁽³⁾.

In breast cancer research, multiple radiomics studies have shown promising results for diagnostic, prognostic, and predictive purposes⁽⁴⁻⁶⁾. Despite these seemingly promising results, translation to clinical practice is limited⁽⁷⁾. A major translational bottleneck can be attributed to the often unknown effect that multiple steps in the radiomics workflow have on feature values, including image acquisition, reconstruction, and pre-processing⁽⁸⁻¹¹⁾. For a radiomics feature to serve as a biomarker, and to be used reliably in clinical decision support systems, it must fulfill the criteria *repeatability* and *reproducibility*⁽¹²⁾. Repeatability can be defined as “the variability of the biomarker when repeated measurements are acquired on the same experimental unit under identical or nearly identical conditions” and reproducibility as to “variability in the biomarker measurements associated with using the imaging instrument in real-world clinical settings, which are subject to a variety of external factors that cannot all be tightly controlled”⁽¹²⁾.

Previous research has already identified several steps in the radiomics workflow that influence the reproducibility and repeatability of radiomics features. For example, image acquisition and reconstruction appear to cause variation in radiomic feature values in research performed on CT imaging^(13, 14). Unlike the Hounsfield Units in CT, MRI does not have absolute signal intensities, potentially causing large differences between images, emphasizing the importance of inspecting and possibly adjusting image intensities before performing feature extraction⁽¹⁵⁾. A test-retest MRI study of glioblastoma showed that both normalization and intensity quantization strategies affect radiomic feature repeatability and that the optimal strategy must be composed per feature group⁽¹⁶⁾. Further test-retest studies assessing feature repeatability have been performed in cervical⁽¹⁷⁾, and prostate cancer^(18, 19) and have shown consistent results, although all studies state that translation of results to other tumor sites has not been confirmed. In contrast, Peerlings et al.⁽²⁰⁾ showed that 9.2% (122/1322) of the features, extracted from apparent diffusion coefficient (ADC) maps in ovarian, liver, and colorectal cancer patients, were

repeatable among the different tumor sites.

The assessment of radiomics feature repeatability by test-retest studies in breast MRI exams is currently lacking. A potential reason for this lack of data is the variance present in a standard clinical breast MRI protocol, which means that scanning parameters may differ between patients scanned with the same clinical protocol. Therefore, this study investigated the repeatability of radiomics feature values extracted from breast MRI exams using a fixed clinical breast protocol comprising of T2-weighted (T2W) images, T1-weighted (T1W) images, and diffusion-weighted images (DWI) and their derived ADC maps.

Material and methods

Study population

The study was approved by the local medical ethical committee and written informed consent was given by all participants before participation. Eleven healthy female volunteers were recruited via college-wide advertisement. Participants were only included if they did not suffer from claustrophobia and met the requirements for admission to the MRI. Participants' height, weight, and the phase of the menstrual cycle were noted. The menstrual cycle of the included healthy volunteers was not taken into account during the MRI exams

Imaging acquisition

All MRI exams were performed using a 16-channel breast coil on one single 1.5 Tesla scanner (Ingenia, Philips Healthcare, Best, The Netherlands) in the same research institution by the same technician. During imaging, the women lay in the prone position with both breasts in the openings of the breast coil and both arms above their head. The performed MRI protocol consisted of a T2-weighted turbo spin echo (T2W), native T1-weighted turbo gradient echo (T1W), and a single shot diffusion-weighted imaging (DWI) sequence using b-values of 0, 150, and 800. A single corresponding ADC-map was derived from all three DWI sequences. All volunteers underwent MRI exams using the identical breast protocol while maintaining as many parameters fixed as possible. The acquisition parameters for the different MRI sequences are shown in the supplementary material (Table S1). The shimbox, needed for the T1W and DWI sequences, was placed on the sternum by default. In case the technician judged the scan as clinically insufficient, the shimbox was placed on the breasts.

Study design

A test-retest study was designed to assess the repeatability of breast-MRI extracted radiomic features. Three separate test-retest strategies were performed twice at six to

ten day intervals. From here on, we will use ‘date 1’ to refer to the first scanning date of each healthy volunteer and ‘date 2’ to refer to the second scanning date. In each strategy, the complete breast MRI protocol was repeated three times with a two-minute pause between each protocol. In the first strategy (S1) the participant remained in the MRI scanner the entire time (including the pauses) without movement, for the acquisition of the three breast MRI protocols. The second strategy (S2) differed from S1 only by moving the table out of the scanner (with the participant still in the same position without movement) during the two-minute breaks. For the third strategy (S3) the participant got off the table during the two minutes breaks (Figure 1). In total, 18 different MRI exams were acquired for each healthy volunteer with a total scanning time of approximately 198 minutes.

ROI segmentation

All images were visually checked for quality (including artifacts) by a dedicated breast radiologist with 14 years of experience (ML) before starting the analysis. The region of interest (ROI) was segmented by a medical researcher (RG) with four years of experience in breast MR imaging and validated by the same dedicated breast radiologist. It was chosen to 3D, manually segment the right breast. The segmentations were bounded by the sternum (medial side), the pectoral muscle (dorsal side), and the axilla (lateral side) in three dimensions using MIM software (version 7.1.3, Cleveland Ohio, United States). Segmentations were performed on all patients on the T2W sequences of all MRI exams as anatomical structures are best visible on this sequence. Subsequently, the T2W sequence was registered with the T1W sequence, and ADC map, using rigid alignments within MIM software, followed by segmentations transfer (Figure 2).

Image pre-processing & feature extraction.

All MRI exams including ROI segmentations were converted to the nearly raw raster data (NRRD) file format using Python (version 3.7.3) for subsequent analysis. Before feature extraction, multiple pre-processing procedures were applied to the images to study their impact on feature repeatability. First, feature extraction was performed without any image pre-processing as a baseline measurement. Second, N4 bias field correction was applied to the images prior to feature extraction⁽²¹⁾. Lastly, the bias field corrected images were further pre-processed using the built-in image z-score normalization by Pyradiomics software (version 2.2.0), with and without binning the voxel grayscale values using a fixed bin width of 32 and 64 (Pyradiomics suggested a bin width between 16-128)^(16, 22). Image pre-processing steps were performed in Python (version 3.7) using an in-house developed pipeline based on the computer vision packages, including OpenCV (version 4.1.0), SimpleITK (version 1.2.0), and NumPy (version 1.16.2). For each ROI, 91 original features were extracted using the Pyradiomics software (version 3.0.1), which is mostly compliant with the Image Biomarker Standardization Initiative

⁽²³⁾. The extracted radiomics feature included first-order statistics features, gray-level co-occurrence matrix features (GLCM), gray-level run length matrix features (GLRLM), gray-level size zone matrix features (GLSZM), neighboring gray tone difference matrix features (NGTDM), and gray-level dependence matrix features (GLDM). All texture features were extracted using default Pyradiomics settings. A detailed Pyradiomics feature description can be found online ⁽²⁴⁾.

Statistical analysis

To assess the repeatability of the extracted radiomic features for the various ROI's in the multiple test-retest strategies, the concordance correlation coefficient (CCC) was calculated using the *epiR* package (Version 0.9-99) (REF) in R language (version 3.6.3) performed in R studio (version 1.2.1335, Vienna Austria) ⁽²⁵⁾. Radiomics features extracted from a given MRI exam are compared to radiomic features extracted from the remaining MRI exams in a pairwise manner. The CCC was used to evaluate the agreement in radiomic feature values, taking into account both the rank and the value of the measurements ⁽²⁶⁾. This metric has the advantage of robust results in small sample sizes ⁽²⁶⁾. The CCC provides values between -1 and 1, with 0 representing no concordance, 1 representing perfect concordance, and -1 perfect inverse concordance. Features with a CCC of > 0.90 were defined as repeatable features, according to suggestions in literature ⁽²⁷⁾. Feature concordance was assessed for each pre-processing procedure using the results of all test-retest strategies of both scanning dates as well as for the results collected on the separate scanning dates. To create an overview of repeatable features across all pairs for the different pre-processing procedures, the intersection of the repeatable features across pairs was noted.

Results

Patients Demographics

The median age of the eleven healthy female volunteers was 28 years (interquartile range 25-30 years). Table 1 summarizes the healthy volunteers' characteristics. Shimbox displacement occurred in 22.6% of the scanned sequences.

Repeatable radiomic features

Due to a scanning error of all T1-weighted images and the ADC maps of one healthy volunteer during scanning date 1, all data of this participant was excluded from the analysis. In both the T1W and T2W sequences as in the ADC maps, in pairwise comparison, the number of concordant features varied per scanning date, per test-retest strategy and, per image pre-processing procedure (Figure 3, 4, and 5). Furthermore, for all pre-processing procedures, the lowest number of concordant features was observed between the MRI exams scanned on date 1 and the MRI exams scanned on date 2, seen

in the reddest field outside the black demarcations in Figures 3, 4 and 5.

TIW Sequence

Across all pairs, regardless of scanning date and test-retest strategy, the highest number of concordant features was seen in the images without pre-processing, resulting in 15/91 (16.5%) concordant features. These 15 features consisted of 7 first-order, 1 GLCM, 2 GLRLM, 2 GLSZM, and 2 GLDM and, 1 NGTDM feature(s) (Table 2). Applying grayscale discretization resulted in 13/91 (14.3%) and 14/91 (15.4%) concordant features for 32-bins and 64-bins, respectively. Compared to the images without pre-processing, the texture features showed less concordant features. The z-score normalized images resulted in the lowest number of 4/91 (4.4%) concordant features. Applying gray-scale discretization after z-score normalization improved the number of concordant textural features to 7/91 (7.7%) and 8/91 (8.8%) for 32-bins and 64-bins, respectively. The loss in the number of concordant features for z-score normalized images (with and without grayscale discretization), when compared to the images without pre-processing, was mainly due to a loss in the number of concordant first-order features, which were 6/91 (6.6%).

For the majority of pre-processing strategies, the images collected during date 2 showed a higher number of concordant features (varying between 10/91 and 48/91 in images without BFC and between 11/91 and 35/91 in BFC images) compared to images collected during date 1 (varying between 4/91 and 32/91 in images without BFC and between 9/91 and 14/91 in BFC images) (Table 3, Figure 3), with these differences being greatest after applying grayscale discretization. Furthermore, for most image pre-processing procedures, the addition of BFC resulted in less concordant features compared to the images without BFC (Table 3, Table S2). For the BFC images without further pre-processing and for the BFC images with grayscale discretization, it was mainly the first-order features that showed a loss of concordance compared to not performing BFC.

Figures S1-S6 present the pairwise CCC values in scatterplots for all features in the different pre-processing procedures, wherein the different colors represent the use of all pairwise comparisons or only the pairwise comparisons between MRI exams scanned on the same day.

T2W Sequence

Across all pairs, regardless of scanning date and test-retest strategy, the z-score normalized images showed the highest number of concordant features, 26/91 (28.6%), of which, 3 first-order, 11 GLCM, 3 GLRLM, 0 GLSZM, 8 GLDM, and 1 NGTDM feature(s) (Table 4). Compared to the other pre-processing procedures, the difference in the number of concordant features was mainly in the concordant texture features, which

were almost non-concordant for the other pre-processing procedures.

The images without pre-processing resulted in 11/91 (12.1%) concordant features across all pairs, of which more than half of these features were first-order features (Table 4). Applying grayscale discretization resulted in a further decrease of concordant features to 7/91 (7.7%) for both 32 and 64 bins. Applying grayscale discretization after z-score normalization resulted in a loss of almost all concordant textural features when compared to z-score normalized images alone. These images resulted in only 4/91 (4.4%) concordant features for both 32 and 64 bins. Notably, the only concordant texture feature (`gldm_SmallDependenceLowGrayLevelEmphasis`) was not concordant after z-score normalization alone.

The addition of BFC resulted in different feature concordance when compared to the same image pre-processing procedures without BFC (Table 4, Table S3). The BFC images without further pre-processing, with 32-bin grayscale discretization and, with 64-bin grayscale discretization resulted in 0/91 (0.0%), 2/91 (2.2%), and 1/91 (1.1%) concordant features, respectively. Despite the overall loss of concordant features, 2/91 (2.2%) features were found to be concordant after the addition of BFC. The BFC z-score normalized images showed the same number of concordant features compared to the z-score normalized images without BFC, although some features improved in concordance, where others lost concordance. The application of grayscale discretization after z-score normalization on BFC images showed the same pattern in concordant features when compared to the images without BFC, namely, a loss of almost all concordant textural features (Table 4 and S3). These pre-processing procedures resulted in 6/91 (6.6%) and 5/91 (5.5%) concordant features, for 32-bins and 64-bins, respectively. Furthermore, it is noteworthy that when looking at the pairwise concordance features for the different scan dates, BFC decreased the feature concordance for MRI exams scanned on date 1, while there was an increase in feature concordance for MRI exams scanned on date 2 (Figure 4, Table 3).

Figures S7-S12 present the pairwise CCC values in scatterplots for all features in the different pre-processing procedures, wherein the different colors represented the use of all pairwise comparisons or only the pairwise comparisons between MRI exams scanned on the same day.

ADC map

Across all pairs, regardless of scanning date and test-retest strategy, the number of concordant features for the images without pre-processing, with 32-bin grayscale discretization, and 64-bin grayscale discretization was 8/91 (8.8%), 7/91 (7.7%), and 6 (6.6%), respectively (Table 5). In none of the pre-processing procedures, first-order

features appeared to be concordant. The number of concordant features was roughly the same for the BFC images with 8/91 (8.8%), 6/91 (6.6%), and 6/91 (6.6%) concordant features for images without further pre-processing, with 32-bin grayscale discretization, and 64-bin grayscale discretization, respectively (Table 5). Although compared to the images without BFC, some features improved in concordance, where others lost concordance (Table 5).

The number of concordant features differed between the images collected on the separated scanning dates, although these differences were minor compared to the T1W and T2W sequences (Figure 5, Table 3). The number of concordant features was 28/91 (30.8%), 15/91 (16.5%) and 11/91 (12.1%) for date 1 and 22/91 (24.1%), 13/91 (14.3%) and 11/91 (12.1%) for date 2, using the images without BFC. The number of concordant features was 9/91 (9.9%), 9/91 (9.9%) and 11/91 (12.1%) for date 1 and 12/91 (13.2%), 12/91 (13.2%) and 11/91 (12.1%) for date 2, using the BFC images.

Figures S13-S15 present the pairwise CCC values in scatterplots for all features in the different pre-processing procedures, wherein the different colors represented the use of all pairwise comparisons or only the pairwise comparisons MRI exams scanned on the same day.

Discussion

In this test-retest study, repeatable radiomics features extracted from breast MRI exams from healthy volunteers were identified, using a fixed scanning protocol including T2-weighted (T2W), unenhanced T1-weighted (T1W), and diffusion-weighted images with corresponding derived ADC maps. This study showed the effects of varying image pre-processing procedures on the radiomics feature repeatability. Across all pairs, the images without pre-processing produced the highest number of repeatable features for both the T1W sequence as well as the ADC maps. In the T2W images, applying z-score normalization produced the highest number of repeatable features.

The assessment of radiomics feature repeatability via test-retest studies in breast MRI exams is currently lacking. The three different MRI sequences examined in this study showed differences in feature repeatability. In addition, the effect of image pre-processing on feature repeatability was different for the two MRI sequences and ADC maps. Not applying image pre-processing produced the highest number of repeatable features in the T1W sequence and the ADC maps. Overall, applying grayscale discretization caused a loss of repeatable textural features in the T1W and T2W sequences, although some texture features became repeatable after grayscale discretization. It is notable that in general, the number of repeatable texture features was reduced after applying

grayscale discretization, although grayscale discretization is considered necessary for the extraction of texture features by both Pyradiomics and the IBSI guidelines⁽²²⁾. Given that MR images do not contain absolute signal values, MRI exams performed on the same scanner using an identical scan protocol could potentially eliminate the need for grayscale discretization. Furthermore, z-score normalized images showed the highest number of repeatable features in the T2W sequence, on the other hand, applying normalization decreased the number of repeatable features in the T1W sequence. Failure to improve the repeatability of features after z-score normalization was also found in the study by Schwier et al.⁽¹⁹⁾, although, in contrast to our results, this was seen in the T2W sequence. They state that image normalization was used to homogenize images acquired from different scanners with different protocols. In our study, however, it was assumed that images scanned with the same protocol on the same scanner were already well comparable in terms of imaging parameters. In addition, the applied normalization uses the whole image for normalization and since the MRI quality decreases further from the coil (at the edges of the images), this reduction in quality can degrade the quality of the breast region (which is close to the coil) and with that the ROI comparability. The same principle could account for the use of BFC since for all sequences it either did not change the number of repeatable features or caused a loss of repeatable features compared to not using BFC. However, failure to improve the repeatability of functions after BFC may also be due to use of default settings for the N4 BFC; findings of Saint Martin et al.⁽²⁸⁾ showed that the default settings for the N4 BFC were not optimal for breast MRI exams.

By considering pairwise comparisons between scans taken on the same day, it was found that for all sequences, including all different preprocessing procedures, except for the T2W sequence and ADC maps without preprocessing, date 2 produced a higher number of repeatable features compared to date 1. One explanation for this may be that the healthy volunteers knew better what to expect on the 2nd scan date after going through the first scan date. In addition, in most cases, the number of repeatable features was higher for the scans taken on the same day compared to the number of repeatable features found from the data of both days, as expected. These differences may be explained by changing factors over time (e.g., system changes in the MRI scanner or biology of the healthy volunteer) that caused variation in the feature values. For example, the homogeneity of the MRI field, gradient systems, and coil affects the image quality⁽²⁹⁾. Furthermore, changes in the biology of the healthy volunteer, including the menstrual cycle and body temperature, are known to affect the MRI exams⁽³⁰⁾. These factors may impact clinical decision making and hence, radiomic features must be robust to these changes.

To date, MRI test-retest studies for the evaluation of repeatable and reproducible features, have been conducted through phantom research^(15, 28, 31-33) and by the use of MRI exams

of healthy volunteers or cancer patients^(17, 19, 20, 32, 34-36). None of these studies investigated feature repeatability and/or reproducibility in human breast MRI exams, and only one study investigated a breast phantom⁽²⁸⁾. The study of Saint Martin et al.⁽²⁸⁾ showed the necessity of image pre-processing dedicated to breast MRI exams before using features in further analysis. Phantom repeatability and reproducibility results seem to be overly optimistic as these overall appear to score higher than the test-retest studies performed within human data. For example, the study by Lee et al.⁽³²⁾ tested feature repeatability in T1W and T2W in both a phantom and MRI brain of healthy volunteers. The average ICC repeatability measures for the T1W and T2W images were higher for the phantom (0.963 and 0.959) compared to healthy volunteers (0.856 and 0.849). Furthermore, a recently published phantom study by Shur et al.⁽³¹⁾ showed that 37/46 (80%) of the radiomic features were concordant ($CCC > 0.9$) in a test-retest study. By contrast, the test-retest study by Eck et al.⁽³⁴⁾ investigating feature repeatability in T2W brain MRI exams of fifteen healthy volunteers showed only 76/146 (52%) of good to excellent repeatable features ($CCC \geq 0.7$). Considering only the excellent repeatable features ($CCC > 0.85$) in the above-mentioned article, the number of concordant features decreased to 40/146 (27.4%), which is more comparable to the results found in this study. The same accounts for a test-retest study in brain MRI exams of glioblastoma patients, in which they identified 386/1043 (37.0%) repeatable features, although they used $CCC > 0.8$ as a cut-off value⁽³⁶⁾. A prostate MRI repeatability study by Schwier et al.⁽¹⁹⁾ concluded that feature repeatability can vary greatly among the radiomic features and that the repeatability of the features is highly sensitive to image pre-processing procedures.

In clinical (prospective) trials, variance in scanners and acquisition and reconstruction parameters between and even within patients is unsurmountable and will therefore affect the reproducibility of the features. Although exploring feature reproducibility was not the aim of this study, this data will be a starting point to investigate the reproducibility of breast MRI extracted radiomic features. Future studies can investigate feature reproducibility by changing the different acquisition parameters one by one while leaving the others fixed. Furthermore, the harmonization method called ComBat, which was originally developed to harmonize gene expression data⁽³⁷⁾, is increasingly being applied in radiomics studies to remove batch effects^(8, 14, 38-40). However, caution should be exercised when applying this harmonization method, as it can only correct for one variable and, MRI data collected from multiple hospitals often contains a multitude of variables. In addition, future studies should focus on the discriminative power of a repeatable and reproducible feature, as a repeatable and reproducible feature does not necessarily imply that this feature is a predictive or prognostic radiomic feature.

Limitations

Firstly, the number of healthy volunteers included was quite limited, although the test-

retest set-up allowed for 18 MRI exams per healthy volunteer, resulting in the analysis of a total of 198 MRI exams. Nevertheless, since this is an early study investigating this topic, we believe that these results are valuable and useful for the radiomics community. Secondly, the included T1W images were examined without adding a contrast agent, so these images cannot be fully compared to the dynamic T1W images normally examined in a clinical breast protocol. Future test-retest studies in breast cancer patients should show whether the repeatable features found in this study are also repeatable in dynamic T1W images. Thirdly, this study investigated only Pyradiomics features extracted from the original image. Future studies could focus more on other feature groups, among others, Gabor, gradient, or Laws. Fourthly, the region of interest contained only healthy tissue, further research in breast cancer patients will have to show whether the repeatable features found in healthy breast tissue can also be considered repeatable in breast tumor tissue. Lastly, it is important to keep in mind that there is a great variety of pre-processing procedures, which can influence feature values. In this study, we choose to use the open-source software Pyradiomics to apply normalization and grayscale discretization to easily reproduce results. In the future, we aim to extend this study with other alternative normalization procedures and focus on feature repeatability.

Conclusions

Varying numbers of repeatable breast MRI radiomic features extracted from healthy volunteers were found for each different test-retest strategy. Furthermore, the effects of image preprocessing procedures on the repeatability of radiomic features were found to be different depending on the MRI sequence.

References

1. Lambin P, Leijenaar RTH, Deist TM, Peerlings J, de Jong EEC, van Timmeren J, et al. Radiomics: the bridge between medical imaging and personalized medicine. *Nat Rev Clin Oncol*. 2017;14(12):749-62.
2. Aerts HJ, Velazquez ER, Leijenaar RT, Parmar C, Grossmann P, Carvalho S, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat Commun*. 2014;5:4006.
3. Ibrahim A, Vallières M, Woodruff H, Primakov S, Beheshti M, Keek S, et al. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Seminars in Nuclear Medicine*. 2019.
4. Liu Z, Li Z, Qu J, Zhang R, Zhou X, Li L, et al. Radiomics of multi-parametric MRI for pretreatment prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer: a multicenter study. *Clin Cancer Res*. 2019.
5. Whitney HM, Taylor NS, Drukker K, Edwards AV, Papaioannou J, Schacht D, et al. Additive Benefit of Radiomics Over Size Alone in the Distinction Between Benign Lesions and Luminal A Cancers on a Large Clinical Breast MRI Dataset. *Acad Radiol*. 2019;26(2):202-9.
6. Bickelhaupt S, Paech D, Kickingereider P, Steudle F, Lederer W, Daniel H, et al. Prediction of malignancy by a radiomic signature from contrast agent-free diffusion MRI in suspicious breast lesions found on screening mammography. *J Magn Reson Imaging*. 2017;46(2):604-16.
7. Conti A, Duggento A, Indovina I, Guerrisi M, Toschi N. Radiomics in breast cancer classification and prediction. *Semin Cancer Biol*. 2020.
8. Ibrahim A, Primakov S, Beuque M, Woodruff HC, Halilaj I, Wu G, et al. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods*. 2020.
9. Tagliafico AS, Piana M, Schenone D, Lai R, Massone AM, Houssami N. Overview of radiomics in breast cancer diagnosis and prognostication. *The Breast*. 2020;49:74-80.
10. Granzier RWY, van Nijnatten TJA, Woodruff HC, Smidt ML, Lobbes MBI. Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review. *European journal of radiology*. 2019;121:108736.
11. Simpson G, Ford JC, Llorente R, Portelance L, Yang F, Mellon EA, et al. Impact of quantization algorithm and number of gray level intensities on variability and repeatability of low field strength magnetic resonance image-based radiomics texture features. *Phys Med*. 2020;80:209-20.
12. Raunig DL, McShane LM, Pennello G, Gatsonis C, Carson PL, Voyvodic JT, et al. Quantitative imaging biomarkers: a review of statistical methods for technical performance assessment. *Stat Methods Med Res*. 2015;24(1):27-67.
13. Varghese BA, Hwang D, Cen SY, Levy J, Liu D, Lau C, et al. Reliability of CT-based texture features: Phantom study. *Journal of Applied Clinical Medical Physics*. 2019;20(8):155-63.
14. Ibrahim A, Refaie T, Primakov S, Barufaldi B, Acciavatti RJ, Granzier RWY, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers*. 2021;13(8).

15. Baessler B, Weiss K, Pinto Dos Santos D. Robustness and Reproducibility of Radiomics in Magnetic Resonance Imaging: A Phantom Study. *Invest Radiol*. 2019;54(4):221-8.
16. Hoebel KV, Patel JB, Beers AL, Chang K, Singh P, Brown JM, et al. Radiomics Repeatability Pitfalls in a Scan-Rescan MRI Study of Glioblastoma. *Radiology: Artificial Intelligence*. 2021;3(1).
17. Fiset S, Welch ML, Weiss J, Pintilie M, Conway JL, Milosevic M, et al. Repeatability and reproducibility of MRI-based radiomic features in cervical cancer. *Radiotherapy and oncology : journal of the European Society for Therapeutic Radiology and Oncology*. 2019;135:107-14.
18. Fedorov A, Vangel MG, Tempany CM, Fennessy FM. Multiparametric Magnetic Resonance Imaging of the Prostate: Repeatability of Volume and Apparent Diffusion Coefficient Quantification. *Invest Radiol*. 2017;52(9):538-46.
19. Schwier M, van Griethuysen J, Vangel MG, Pieper S, Peled S, Tempany C, et al. Repeatability of Multiparametric Prostate MRI Radiomics Features. *Sci Rep*. 2019;9(1):9441.
20. Peerlings J, Woodruff HC, Winfield JM, Ibrahim A, Van Beers BE, Heerschap A, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep*. 2019;9(1):4800.
21. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010;29(6):1310-20.
22. van Griethuysen JJM, Fedorov A, Parmar C, Hosny A, Aucoin N, Narayan V, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res*. 2017;77(21):e104-e7.
23. Zwanenburg A, Vallières M, Abdalah MA, Aerts HJWL, Andrearczyk V, Apte A, et al. The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. *Radiology*. 2020.
24. Pyradiomics feature description [Available from: <https://pyradiomics.readthedocs.io/en/latest/features.html>].
25. Racine JS. RStudio: a platform-independent IDE for R and Sweave. *Journal of Applied Econometrics*. 2012;27(1):167-72.
26. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989;45(1):255-68.
27. McBride G. A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. NIWA Client Report: HAM2005-062. 2005.
28. Saint MJS, Orhac F, Akl P, Khalid F, Nioche C, Buvat I, et al. A radiomics pipeline dedicated to Breast MRI: validation on a multi-scanner phantom study. *Magn Reson Mater Phys*. 2021;34(3):355-66.
29. Jackson E. MR Acceptance Testing and Quality Control: Report of AAPM MR Subcommittee TG1. *Med Phys Journal Translated Name Medical Physics*. 2009;36(6).
30. Dontchos BN, Rahbar H, Partridge SC, Lehman CD, DeMartini WB. Influence of Menstrual Cycle Timing on Screening Breast MRI Background Parenchymal Enhancement and Diagnostic Performance in Premenopausal Women. *J Breast Imaging*. 2019;1(3):205-11.
31. Shur J, Blackledge M, D'Arcy J, Collins DJ, Bali M, O'Leach M, et al. MRI texture feature repeatability and image acquisition factor robustness, a phantom study and in silico study. *Eur*

Test-retest data for the assessment of breast MRI radiomic feature repeatability

- Radiol Exp. 2021;5(1):2.
32. Lee J, Steinmann A, Ding Y, Lee H, Owens C, Wang J, et al. Radiomics feature robustness as measured using an MRI phantom. *Sci Rep.* 2021;11(1):3973.
 33. Dreher C, Kuder TA, Konig F, Mlynarska-Bujny A, Tenconi C, Paech D, et al. Radiomics in diffusion data: a test-retest, inter- and intra-reader DWI phantom study. *Clin Radiol.* 2020;75(10):798 e13-e22.
 34. Eck B, Chirra PV, Muchhala A, Hall S, Bera K, Tiwari P, et al. Prospective Evaluation of Repeatability and Robustness of Radiomic Descriptors in Healthy Brain Tissue Regions In Vivo Across Systematic Variations in T2-Weighted Magnetic Resonance Imaging Acquisition Parameters. *J Magn Reson Imaging.* 2021.
 35. Moradmand H, Aghamiri SMR, Ghaderi R. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *J Appl Clin Med Phys.* 2020;21(1):179-90.
 36. Kickingreder P, Neuberger U, Bonekamp D, Piechotta PL, Gotz M, Wick A, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro Oncol.* 2018;20(6):848-57.
 37. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-27.
 38. Fortin J-P, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage.* 2017;161:149-70.
 39. Fortin J-P, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage.* 2018;167:104-20.
 40. Ibrahim A, Refaee T, Leijenaar RTH, Primakov S, Hustinx R, Mottaghy FM, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One.* 2021;16(5):e0251147.

Chapter 10

Table I: Patient characteristics.

	healthy volunteers (n=11)
Age (years) (median; IQR)	28 (25 - 30)
Height (cm) (median; IQR)	167 (167 - 172)
Weight (kg) (median; IQR)	60 (58 - 63)
Week of the menstrual cycle*	Date 1 / Date 2
Week 1	1 / 5
Week 2	1 / 1
Week 3	3 / 1
Week 4	4 / 2
Days between scan (mean; range)	7 (6 - 9)

* no measurement of the menstrual cycle possible for two healthy volunteers.

Abbreviations: IQR, Interquartile range.

Table 2: Concordant features across all pairs for the T1-weighted MRI exams, with A: no pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization.

	A	B	C	D	E	F
Number of concordant features	15	13	14	4	7	8
	(16.5%)	(14.3%)	(15.4%)	(4.4%)	(7.7%)	(8.8%)
firstorder_90Percentile	x	x	x			
firstorder_InterquartileRange	x	x	x			
firstorder_MeanAbsoluteDeviation	x	x	x			
firstorder_Mean	x	x	x			
firstorder_RobustMeanAbsoluteDeviation	x	x	x			
firstorder_RootMeanSquared	x	x	x			
firstorder_Skewness	x	x	x	x	x	x
glcm_JointAverage	x					
gldm_GrayLevelNonUniformity	x	x	x		x	x
gldm_RunLengthNonUniformity	x	x	x		x	x
glszm_GrayLevelNonUniformity	x		x	x		x
glszm_SizeZoneNonUniformity				x		
glszm_SmallAreaHighGrayLevelEmphasis	x					
gldm_DependenceNonUniformity	x	x	x		x	x
gldm_GrayLevelNonUniformity	x	x	x	x	x	x
ngtdm_Busyness		x	x		x	x
ngtdm_Coarseness	x	x	x		x	x

Test-retest data for the assessment of breast MRI radiomic feature repeatability

Table 3: Number of concordant features across all pairs for the entire dataset (All) and across all pairs from the separate scanning dates (Date 1 and Date 2) for all sequences with and without bias field correction, with A: no further pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization.

Sequences	Without BFC			With BFC		
	All	Date 1	Date 2	All	Date 1	Date 2
T1W						
A	15	32	40	8	13	11
B	13	19	45	10	11	30
C	14	18	48	8	12	31
D	4	4	10	4	9	12
E	7	10	35	10	13	34
F	8	9	38	8	14	35
T2W						
A	11	31	16	0	1	60
B	7	9	12	2	3	22
C	7	9	11	1	3	23
D	26	35	44	26	39	37
E	4	7	7	6	11	17
F	4	7	6	5	11	18
ADC						
A	8	28	22	8	9	12
B	7	15	13	6	9	12
C	6	11	11	6	11	11

Chapter 10

Table 4: Concordant features across all pairs for the T2-weighted MRI exams, with A: no pre-processing, B: 32-bin grayscale discretization, C: 64-bin grayscale discretization, D: Z-score normalization, E: Z-score normalization + 32-bin grayscale discretization, and F: Z-score normalization + 64-bin grayscale discretization.

	A	B	C	D	E	F
Number of concordant features	11 (12.1%)	7 (7.7%)	7 (7.7%)	26 (28.6%)	4 (4.4%)	4 (4.4%)
firstorder_10Percentile				x	x	x
firstorder_90Percentile	x	x	x			
firstorder_InterquartileRange	x	x	x	x	x	x
firstorder_MeanAbsoluteDeviation	x	x	x			
firstorder_Mean	x	x	x			
firstorder_RobustMeanAbsoluteDeviation	x	x	x	x	x	x
firstorder_RootMeanSquared	x	x	x			
glcm_JointAverage	x					
glcm_Contrast				x		
glcm_DifferenceAverage	x			x		
glcm_DifferenceEntropy				x		
glcm_DifferenceVariance				x		
glcm_JointEntropy				x		
glcm_Idm				x		
glcm_Idmn				x		
glcm_Id				x		
glcm_Idn				x		
glcm_InverseVariance				x		
glcm_SumEntropy				x		
gldm_GrayLevelNonUniformity				x		
gldm_RunLengthNonUniformity	x					
gldm_RunPercentage				x		
gldm_RunVariance				x		
gldm_DependenceEntropy				x		
gldm_DependenceNonUniformity				x		
gldm_DependenceNonUniformityNormalized				x		
gldm_DependenceVariance				x		
gldm_GrayLevelNonUniformity				x		
gldm_LargeDependenceEmphasis				x		
gldm_LargeDependenceHighGrayLevelEmphasis				x		
gldm_SmallDependenceHighGrayLevelEmphasis				x		
gldm_SmallDependenceLowGrayLevelEmphasis	x	x	x		x	x
ngtdm_Complexity				x		
ngtdm_Contrast	x					

Test-retest data for the assessment of breast MRI radiomic feature repeatability

Table 5: Concordant features across all pairs for the ADC maps, with A: no pre-processing, B: 32-bin grayscale discretization, and C: 64-bin grayscale discretization, D: bias field correction, E: bias field correction + 32-bin grayscale discretization and, F: bias field correction + 64-bin grayscale discretization.

	A	B	C	D	E	F
Number of concordant features	8 (8.8%)	7 (7.7%)	6 (6.6%)	8 (8.8%)	6 (6.6%)	6 (6.6%)
glcm_ClusterProminence	x					
glcm_Correlation	x	x	x	x	x	x
glcm_Imc1	x	x	x	x		x
glcm_Imc2	x	x	x	x	x	x
gldm_GrayLevelNonUniformity		x	x		x	x
gldm_RunLengthNonUniformity	x	x	x	x	x	x
glszm_GrayLevelNonUniformity	x	x	x	x	x	x
glszm_SizeZoneNonUniformity	x			x		
gldm_DependenceNonUniformity	x			x		
ngtdm_Coarseness		x		x	x	

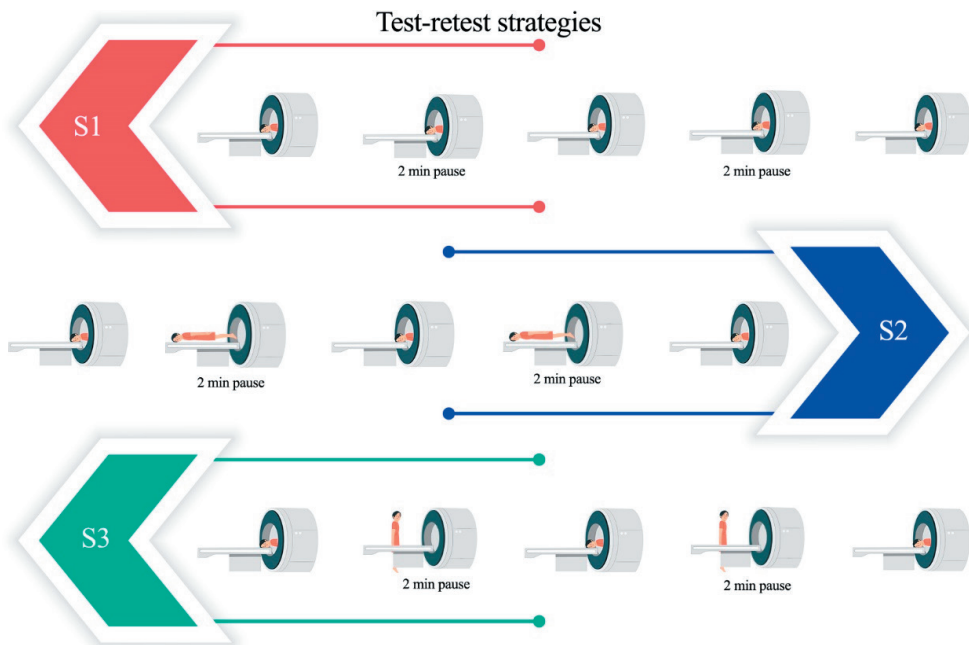


Figure I: Visual representation of the three test-retest strategies.



Figure 2: An axial slice of a 3D MRI exam of a healthy volunteer including right breast segmentation (red margin). A: ADC map, B: T2-weighted image, C: T1-weighted image.



111

Chapter 11

MaasPenn radiomics
reproducibility score:
a novel quantitative measure for
evaluating the reproducibility of
CT-based hand-crafted radiomic features

Authors

Abdalla Ibrahim, Bruno Barufaldi, Turkey Refae, Telmo M. Silva Filho,
Raymond J. Acciavatti, Zohaib Salahuddin, Roland Hustinx, Felix M. Mottaghy,
Andrew D.A. Maidment and Philippe Lambin

Adapted from

Cancers. Submitted.

Abstract

The reproducibility of handcrafted radiomic features (HRFs) has been reported to be affected by variations in imaging parameters, which significantly affect the generalizability of developed signatures and translation to clinical practice. However, the collective effect of the variations in imaging parameters on the reproducibility of HRFs remains unclear, with no objective measure to assess it in the absence of reproducibility analysis. We assessed these effects of variations in a large number of scenarios, and developed the first quantitative score to assess the reproducibility of CT-based HRFs without the need for phantom or reproducibility studies. We further assessed the potential of image resampling and ComBat harmonization for removing these effects. Our findings suggest the need for radiomics-specific harmonization methods, and our developed score will serve as a guide to develop generalizable radiomic signatures and ease its incorporation in clinical practice.

Keywords:

Radiomics reproducibility, Image interpolation, ComBat harmonization, CT radiomics

Introduction

Recent decades witnessed a leap in the development of medical imaging and computational powers. Combined with the advancement in artificial intelligence and its inclusion in various activities, an opportunity for converting medical images into mineable quantitative data was created, and the field of radiomics emerged as a result ¹. Handcrafted radiomics -‘the high throughput extraction of mineable quantitative features from medical imaging’ ²- gained exponential research momentum within the last decade. Radiomics offer an alternative for invasive procedures for clinical diagnosis, and could play a significant role in early detection and personalized management ³. Due to the heterogeneity of tumors ^{4,5}, clinical approaches, such as tissue biopsies, might fail to characterize the entirety of the tumor, or require another trial ⁶. In contrast, the radiomics approach takes the regions of interest (ROIs) as input, which could allow better characterization of the lesion ⁷. Moreover, radiomic signatures could offer relatively fast, non-invasive, highly accurate, and cost-effective clinical biomarkers, which will ultimately improve personalized care.

Handcrafted radiomic features (HRFs) can decode biological information from suspicious tissue under study ³ as potential clinical biomarkers. To date, many studies reported on the potential of HRFs to predict clinical endpoints, such as detection and diagnosis, response to treatment, overall survival and progression free survival. On the other hand, a number of limitations that hinder the clinical translation of the developed radiomic signatures have been identified. The mainstay of a biomarker is the ability to quantify it in a reproducible manner ⁸. As HRFs are calculated using data-characterisation algorithms applied to the medical image, changes in scan acquisition and reconstruction parameters can significantly affect the reproducibility of HRFs. A fraction of HRFs has been reported to be sensitive to variations in the acquisition and reconstruction parameters of the scans, and the number of reproducible HRFs is usually dependent on the degree of variation in these parameters ⁹⁻¹³.

A number of studies investigated the potential of harmonization methods, such as ComBat, to remove variations attributed to differences in acquisition and reconstruction parameters ¹⁴⁻¹⁷. ComBat harmonization was originally developed to harmonize gene expression arrays ¹⁸, and has shown promising results in radiomics analyses in certain scenarios ¹⁴⁻¹⁷. However, there is no consensus on how or when to use ComBat harmonization in radiomics.

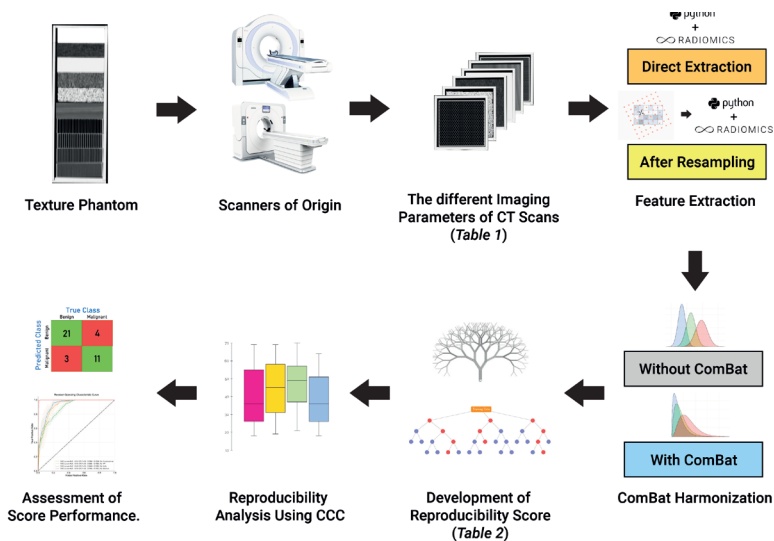
We previously published a framework to assess the reproducibility of radiomic features ⁷, with two follow up studies to validate it on a phantom dataset ^{9,11}. A number of studies investigated the effects of different parameters individually on the reproducibility of

HRFs ^{19,20}. However, the collective effect of variations in more than a single imaging parameter at a time is yet to be investigated. Furthermore, there is currently no quantitative measure to evaluate the reproducibility of HRFs in a given dataset. In this study, we investigated the effect of variations in imaging parameters on different imaging scenarios of phantom scans. We aimed to develop an objective metric to assess the reproducibility of HRFs across scans, which could be used as an indicator to assess the data under analysis, and further as a tool to ‘quality check’ radiomic studies.

Methods

Imaging data

The publicly available Credence Cartridge Radiomics phantom dataset ²¹ was analysed in this study (available on: TCIA.org) ²². The dataset consists of 251 scans of a phantom that were acquired with different imaging vendors, models, and imaging parameters. The workflow applied in this study is shown in figure 1.



Manufacturer	Manufacturer Model	Convolution Kernel	Slice thickness (mm)	Pixel spacing (mm)	
GE MEDICAL SYSTEMS	Discovery STE	STANDARD	1.25, 2.5, 3.75	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		LIGHT_SPEED	1.25, 2.5, 3.75	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		PRO 32	1.25, 2.5, 3.75	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
Philips	Brilliance 64	B.L.C.A.YA	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		B	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		Brightness Big Bore	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
SIEMENS	SOMATOM Definition AS	I241, I301, I311, I401, I441, I521, I701	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		Sensation 16	I311a	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98
		Sensation 40	B10F, B20F, B30F, B11c, B21c, B31c, B40F, B70F	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98
SIEMENS	Sensation 64	B10F, B20F, B30F, B11c, B21c, B31c, B40F, B70F	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98	
		Sensation 40	B10F, B20F, B30F, B11c, B21c, B31c, B40F, B70F	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98
		Sensation 64	B10F, B20F, B30F, B11c, B21c, B31c, B40F, B70F	1.5, 2, 3	0.39, 0.49, 0.59, 0.68, 0.78, 0.88, 0.98

Table 1 : The different imaging parameters of the phantom CT scans.

Manufacturer	Manufacturer Model Name	Convolution Kernel	Value assigned
SIEMENS	Sensation 40, Sensation 64	B10F	0.162
GE MEDICAL SYSTEMS	Discovery STE	SOFT	0.163
Philips	Brilliance 64	A	0.167
Philips	Brilliance 64	Y6	0.200
SIEMENS	Sensation 40, Sensation 64	B20F	0.284
Philips	Brilliance 64, Brilliance Big Bore	B	0.333
SIEMENS	SOMATOM Definition AS	I241	0.376
SIEMENS	Sensation 16	B30F	0.426
SIEMENS	SOMATOM Definition AS	I301	0.428
GE MEDICAL SYSTEMS	Discovery STE, LightSpeed Pro 32	STANDARD	0.429
SIEMENS, Philips	Sensation 16, Sensation 40, Sensation 64, Brilliance 64	B21c, C	0.433
SIEMENS	SOMATOM Definition AS	I311a	0.435
SIEMENS	Sensation 40, Sensation 64	B31F	0.440
GE MEDICAL SYSTEMS	Discovery STE	DETAIL	0.443
GE MEDICAL SYSTEMS	Discovery STE	EDGE	0.571
SIEMENS	SOMATOM Definition AS	I401	0.574
SIEMENS	SOMATOM Definition AS	I441	0.601
Philips	Brilliance 64	L	0.667
SIEMENS	Sensation 40, Sensation 64	B50F	0.709
GE MEDICAL SYSTEMS	Discovery STE	LUNG	0.716
SIEMENS	SOMATOM Definition AS	I501	0.716
SIEMENS	Sensation 40, Sensation 64	B60F	0.851
SIEMENS	Sensation 40, Sensation 64	B70F	0.993
SIEMENS	SOMATOM Definition AS	I701	1.002

Table 2 : Numeric values assigned to convolution kernels.

Figure 1. Explanatory diagram of the workflow applied.

Volumes of interest and HRF extraction

Each layer of the phantom (in total 10 layers) was subdivided into 16 equal volumes of interest (VOI), sized $2 \times 2 \times 2 \text{ cm}^3$. A total of 160 VOIs were segmented per scan, resulting in a total of 40160 analysed VOIs. HRFs were extracted using the open source PyRadiomics software²³. HRFs were extracted three different times: (i) directly from the original scans; (ii) from following resampling of all scans to the median resolution available in the dataset; (iii) from following resampling of all scans to the lowest resolution available in the dataset. Image intensities were binned in all of the three scenarios with a binwidth of 25 Hounsfield Units (HUs) to reduce noise levels and texture matrix sizes, and therewith the required computational power. No further image preprocessing was applied in (i)-(iii). Extracted HRFs included HU intensity features, and texture features that describe the spatial distribution of voxel intensities using five matrices: (i) grey-level co-occurrence (GLCM); (ii) grey-level run-length (GLRLM); (iii) grey-level size-zone (GLSZM); (iv) grey-level dependence (GLDM); and (v) neighborhood grey-tone difference (NGTDM) matrices. A more detailed description of PyRadiomics HRFs can be found online at <https://pyradiomics.readthedocs.io/en/latest/features.html>.

Exploratory analysis

All statistical analyses were performed using the R language²⁴ on RStudio (V 3.6.3)²⁵. We performed an initial exploratory analysis to assess the reproducibility of HRFs in the different scenarios mentioned above, as well as the use of ComBat harmonization¹⁸ and Cosine Windowed Sinc (CWS) image interpolation²⁶. The concordance correlation coefficient (CCC) was used to assess the reproducibility of HRFs across the different pairwise scenarios²⁷, using epiR package²⁸. The CCC measures the concordance in both value and rank in each of the pairwise scenarios. HRFs with $\text{CCC} > 0.9$ were considered reproducible. Further details and results are presented in the supplementary materials. The reproducibility of HRFs was assessed in: (i) HRFs extracted from the original scans, before and after ComBat harmonization; (ii) HRFs extracted from scans resampled to the median voxel size ($0.68 \times 0.68 \times 1.5 \text{ mm}^3$), before and after ComBat harmonization; and (iii) HRFs extracted from scans resampled to the largest voxel size ($0.98 \times 0.98 \times 3.75 \text{ mm}^3$), before and after ComBat harmonization.

Evaluation of the effects of variations in imaging parameters

To unravel the effects of variations in imaging parameters, we assessed the reproducibility of HRFs across each pair of the 251 scans, resulting in a total of 31375 pairs (scenarios) analysed. Each of the parameters: (i) Vendor; (ii) Model; (iii) Tube Current; (iv) Exposure; (v) Exposure time; (vi) Slice thickness; (vii) Pixel spacing; and (viii) Convolution kernel was given a numeric value between 0 and 1 depending on the scenario. For vendor and model, we assigned a binary value of 0 -in case the vendor/model is different across the pairs, and 1- in case the same vendor/model was used to acquire both scans in

the scenario. For the remaining parameters, a value between 0 and 1 was calculated by dividing the minimum value of a given parameter by the maximum value across the pairs being analysed. Convolution kernels were assigned a numeric value based on #### (Table 2). To assess the impact, as well as the predictive power, of the variations in imaging parameters on the percentage of reproducible HRFs in different scenarios, a random forest model ²⁹ was applied.

Quantitative score development

After training a random forest on the 31375 pairs, the parameters with the largest feature importance in the model were used to develop a quantitative score. The parameters for the random forest were....###. The selected parameters were multiplied by their importance and divided by the total importance of the included parameters. The sum of weighted parameters was used as a quantitative score with values ranging between ~0.3 and 1. The correlation of the developed score with the percentage of reproducible HRFs across the investigated scenarios was assessed using spearman correlation ³⁰.

To develop a methodology for applying the developed score in radiomic studies, we used different thresholds (increments of 10% between 10% and 90%) of the percentage of reproducible features across the scenarios. The thresholds were used to create a binary label for the percentage of reproducible HRFs in a given scenario, where 0 indicated that the number of reproducible HRFs was below the threshold, and 1 indicated that the percentage was higher than the threshold. Another random forest model was then trained using the binary status of pairs as the outcome. The performance of the cut-off point score was assessed for each of the thresholds defined.

To assess the robustness of the quantitative score, the analysis was repeated 100 times, and the scenarios (pairs) were split randomly into 70% training and 30% validation in each of the runs. Area under the receiver operator characteristics curve (AUC) ³¹, sensitivity and specificity ³² were used to assess the performance of the developed score in predicting whether the percentage of reproducible HRFs in a given scenario was above the selected threshold.

To identify HRFs that were insensitive to variations in imaging parameters, the intersection of reproducible HRFs across all the scenarios was obtained. Similarly, HRFs that were harmonizable using ComBat harmonization ¹⁸ and/or CWS interpolation were identified by obtaining the intersection of HRFs that were found to be reproducible across all pairs following the application of a given harmonization method.

Results

The reproducibility of HRFs across pairs

The number (percentage) of reproducible HRFs extracted directly from the original images varied depending on the differences in imaging parameters across each of the analysed pairs, with a mean of 25.6 (28.1%) HRFs and a standard deviation of 14.4. The average numbers of reproducible HRFs following image resampling to the median and lowest resolutions were 29 (31.9%) +/- 16.6, and 26 (28.6%) +/- 15.5.

Reproducible and Harmonizable HRFs

We identified four HRFs that were insensitive to all variations in the investigated 31375 scenarios. These HRFs are: (i) original first order mean; (ii) original first order median; (iii) original first order root mean squared; and (iv) original first order total energy. One additional HRF (original first order energy) was found to be reproducible across all scenarios following image resampling both to the median and to the largest voxel size available using CWS interpolation. Similarly, one additional HRF was found to be reproducible across all scenarios following the application of ComBat harmonization on HRFs extracted from original scans (original first order 10 percentile), or from scans after resampling to the largest voxel size available (original first order energy). Two additional HRFs (original first order 10 percentile and original first order energy) were found to be reproducible across all pairs following the application of ComBat harmonization on HRFs extracted following resampling to the median voxel size available. The reproducibility and harmonizability (using ComBat or image resampling) of the remaining HRFs were dependent on the variations in imaging parameters across the pairs being analysed. On average, ComBat harmonization outperformed image resampling. The distributions of the percentages of reproducible features in all of the investigated scenarios are shown in Figure 2.

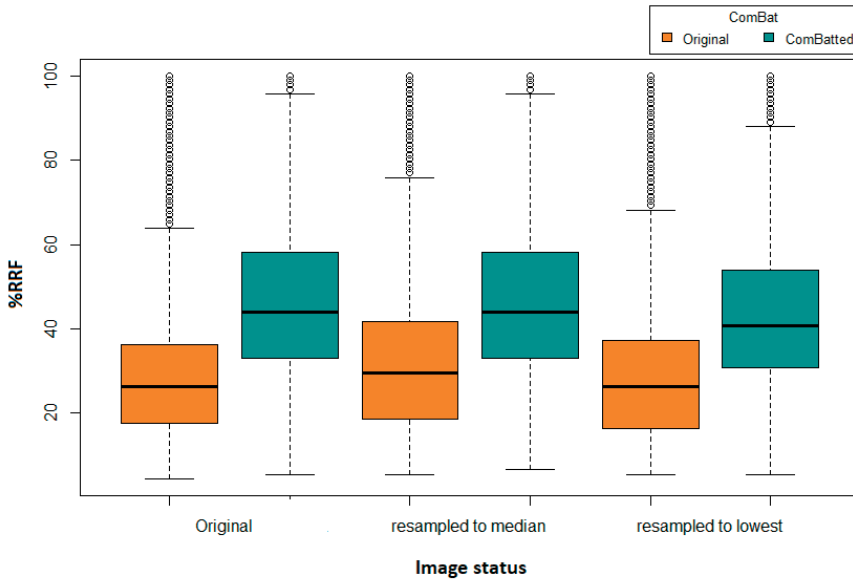


Figure 2. Boxplot of the number of reproducible HRFs in different scenarios.

The effects of variations in imaging parameters

Differences in convolution kernels were found to be the most important factor affecting the reproducibility of HRFs across CT scans acquired differently. The second most important factor was found to be the differences in slice thickness, followed by differences in pixel spacing. The initial random forest was able to explain 82.5% of the variance in the percentage of reproducible HRFs in all of the scenarios investigated.

MPenn radiomics reproducibility score

Based on the random forest model, the difference in convolution kernel had the highest contribution to the score with 48% of the total score. The differences in slice thickness and pixel spacing corresponded to 33% and 19%, respectively. If the scans were acquired with the same (or similar) convolution kernel, the same slice thickness, and pixel spacing (MPenn score >0.98), then the probability of having 90% or more of the HRFs reproducible is 0.97, with a 3% false alarm rate. In contrast, the probability of having 10% or less reproducible HRFs across scans acquired with different convolution kernels and voxel sizes (MPenn score <0.75) is 0.74, and a 19% false alarm rate. The predictive power of our developed Maastricht-Pennsylvania Radiomics Reproducibility Score (MPenn radiomics reproducibility score) to determine the percentages (thresholds) of reproducible HRFs across scans acquired differently is reported in Table 1.

Table I. Performance of the score threshold for the identification of different HRFs reproducibility thresholds.

Percentage RRFs	Score	AUC	CI95% lower	CI95% upper	Specificity	Sensitivity	False alarm
Threshold 10 %	0.75	0.86	0.855	0.867	0.81	0.74	0.19
Threshold 20 %	0.77	0.85	0.842	0.851	0.76	0.77	0.24
Threshold 25 %	0.80	0.85	0.843	0.852	0.80	0.74	0.20
Threshold 30 %	0.83	0.86	0.851	0.86	0.84	0.73	0.16
Threshold 40 %	0.85	0.87	0.868	0.878	0.81	0.80	0.19
Threshold 50 %	0.88	0.90	0.892	0.904	0.83	0.85	0.17
Threshold 60 %	0.88	0.92	0.91	0.925	0.79	0.92	0.21
Threshold 70 %	0.94	0.96	0.952	0.966	0.94	0.89	0.06
Threshold 75 %	0.95	0.97	0.967	0.977	0.95	0.93	0.05
Threshold 80 %	0.96	0.98	0.971	0.983	0.95	0.95	0.05
Threshold 90 %	0.98	0.99	0.982	0.996	0.97	0.97	0.03

Robustness of MPenn radiomics reproducibility score

The confirmatory analysis of the robustness of the MPenn radiomics score was based on the experiment with 100 runs. The results showed a narrow distribution of values across the different metrics with similar performances in the training and validation sets. Figure 2 shows the distributions of AUC values on the training and validation datasets across the 100 runs.

Chapter II

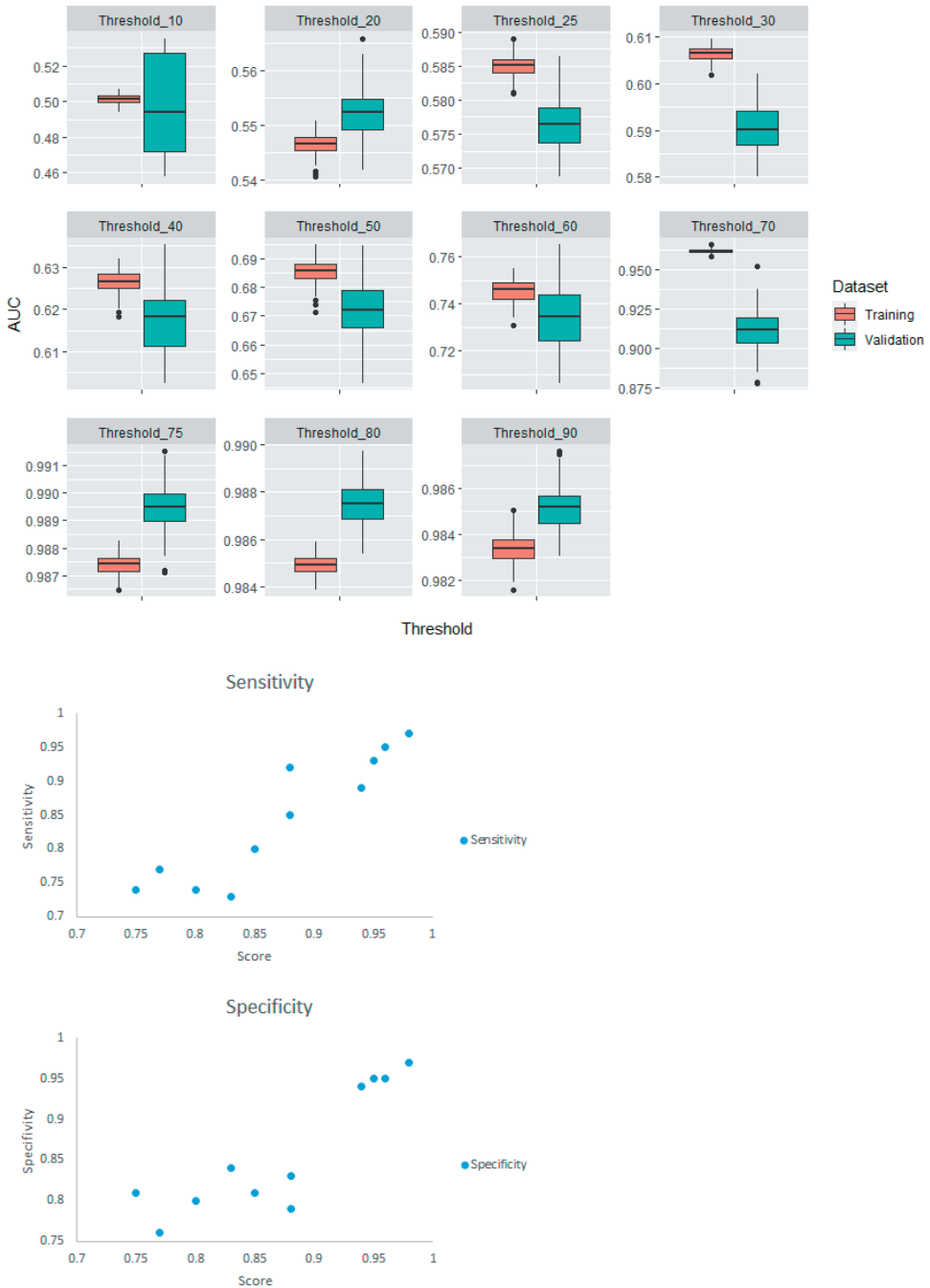


Figure 3. (a) AUC distributions across 100-runs for MPenn radiomics reproducibility score in the training and validation datasets for each of the thresholds of percentage reproducible HRFs; (b) The sensitivity as a function of the score.

Discussion

In this study, we aimed to investigate the effects of variations in CT imaging parameters on the reproducibility of HRFs on a phantom dataset. The scans (n=251) were acquired using a wide range of imaging parameters on different imaging vendors and models. The imaging parameters could be classified into three groups: (i) resolution parameters: convolution kernel, slice thickness and pixel spacing; (ii) noise parameters: mAs, exposure and exposure time; and (iii) hardware make: vendor and model. Our analysis showed that variations in resolution parameters had the most pronounced effects on the reproducibility of HRFs, with the differences in convolution kernel being the most significant contributor. Scans acquired with the same or similar convolution kernels showed the highest numbers of reproducible HRFs across scenarios. Slice thickness and pixel spacing were the other major contributors to the reproducibility of CT based HRFs. A previous study that investigated the reproducibility of HRFs on lung CT scans reconstructed using two different kernels reported that HRFs extracted from scans reconstructed with these two kernels should not be used interchangeably¹³, which is in line with our findings. An important finding in this study is that differences in imaging vendor and model did not seem to affect the reproducibility of HRFs significantly, given that the remaining parameters were similar/homogenous.

We further identified the HRFs that were reproducible regardless of the variations in imaging parameters in our dataset. These were strictly first order features that are descriptive of the HUs in the defined VOIs. This finding can be supported by the fact that HUs are standardized. Henceforth, HRFs such as mean or median HU value are expected to be reproducible across all imaging variations. Lu et al investigated the reproducibility of HRFs by reconstructing raw CT scans of 32 lung cancer patients using different imaging parameters, which ultimately resulted in 15 different scenarios¹². The authors reported that 23/89 (25.8%) HRFs were found to be reproducible across their investigated scenarios, which is also in concordance with our finding that on average, ~26/91 (28.1%) of the HRFs were found to be reproducible across all investigated scenarios. In addition, we identified HRFs that can be harmonized with ComBat or CWS image resampling regardless of the variations in imaging parameters across the scans being analyzed. Both methods could harmonize 1% additional HRFs, and the combination of ComBat harmonization and resampling to median voxel size resulted in an additional 2% of the HRFs across all scenarios. The ability of both methods to harmonize the remaining HRFs was dependent on the variations in imaging parameters in the scenarios analysed. These findings are in line with our previous experiments, which also showed that the reproducibility and harmonizability of the majority of HRFs are dependent on the variations in imaging parameters^{9-11,33}.

In addition, we have successfully developed a quantitative score (MPenn radiomics reproducibility score), which can estimate the percentages of reproducible HRFs across CT scans acquired differently. MPenn radiomics score is the first quantitative tool for assessing the reproducibility of CT based HRFs. It can serve as a screening tool for the inclusion of CT scans in a dataset/study. We performed an extra analysis to assess the robustness of our developed score. The results showed very narrow distributions of performance metric values that were consistent on the training and validation sets across the 100 random splits, which suggests that MPenn radiomics reproducibility score is robust.

While the phantom dataset analysed included a large number of scans acquired with a wide variety of imaging vendors and parameters, a number of the CT imaging vendors used in some clinics were not available for this study. As such, and despite the large number of scenarios investigated, the generalizability of MPenn radiomics score to CT scans acquired with those imaging vendors/parameters has to be investigated. Furthermore, while the phantom used in this study was designed specifically for radiomics, it might not reflect the exact situation of real patients. Future studies that include cadaveric/3D-printed tissues scanned with a larger number of imaging vendors/parameters could better represent patient scans, and can further enhance the utility of the MPenn radiomics score.

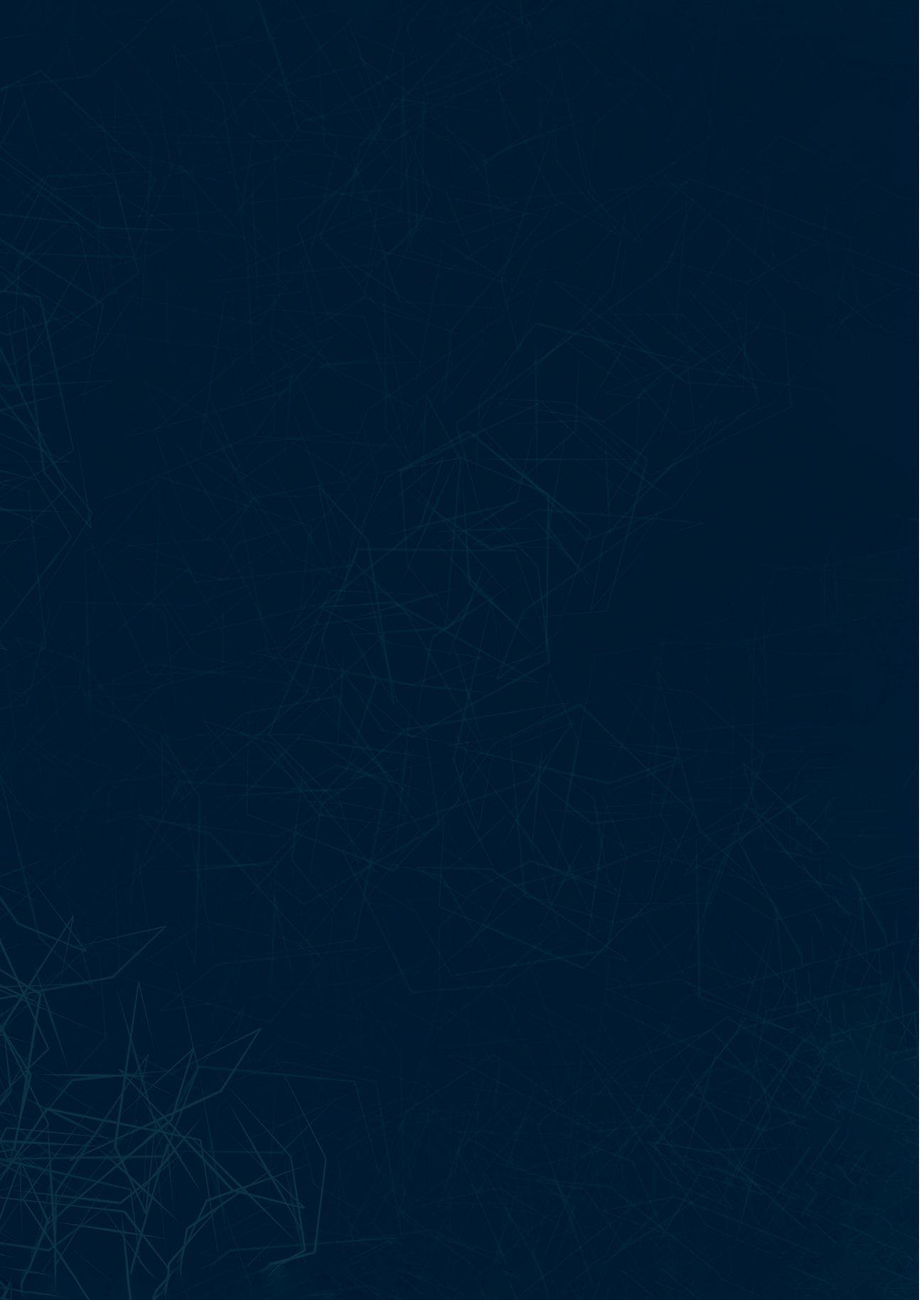
In conclusion, convolution kernel and voxel size differences significantly affect the reproducibility of HRFs. Harmonization methods, such as image resampling and ComBat harmonization, can increase the number of reproducible HRFs with varying degrees, depending on the variations in the imaging parameter of the scans being analyzed. Most significantly, we developed the MPenn score, which can be used to predict the percentage of reproducible HRFs across CT scans acquired differently. Further research with a larger number of scans of cadaveric/3D-printed tissues can further improve the predictive power of the MPenn radiomics score. The development of HRFs that are insensitive to variations in imaging parameters is another potential solution for developing generalizable radiomic signatures.

References

1. Walsh S, de Jong EEC, van Timmeren JE, et al. Decision Support Systems in Oncology. *JCO Clin Cancer Inform.* 2019;3:1-9. doi:10.1200/CCI.18.00001
2. Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. *Eur J Cancer.* 2012;48(4):441-446. doi:10.1016/j.ejca.2011.11.036
3. Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images Are More than Pictures, They Are Data. *Radiology.* 2016;278(2):563-577. doi:10.1148/radiol.2015151169
4. Swanton C. Intratumor heterogeneity: evolution through space and time. *Cancer Res.* 2012;72(19):4875-4882. doi:10.1158/0008-5472.CAN-12-2217
5. Gerlinger M, Rowan AJ, Horswell S, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med.* 2012;366(10):883-892. doi:10.1056/NEJMoa1113205
6. Soo TM, Bernstein M, Provias J, Tasker R, Lozano A, Guha A. Failed stereotactic biopsy in a series of 518 cases. *Stereotact Funct Neurosurg.* 1995;64(4):183-196. doi:10.1159/000098747
7. Ibrahim A, Primakov S, Beuque M, et al. Radiomics for precision medicine: current challenges, future prospects, and the proposal of a new framework. *Methods.* Published online June 3, 2020. doi:10.1016/j.ymeth.2020.05.022
8. Strimbu K, Tavel JA. What are biomarkers? *Curr Opin HIV AIDS.* 2010;5(6):463. <https://www.ncbi.nlm.nih.gov/pmc/articles/pmc3078627/>
9. Ibrahim A, Refaee T, Primakov S, et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers .* 2021;13(8). doi:10.3390/cancers13081848
10. Ibrahim A, Primakov S, Barufaldi B, et al. Reply to Orlhac, F; Buvat, I. Comment on "Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* 2021, 13, 1848." *Cancers .* 2021;13(12):3080. <https://www.mdpi.com/1157110>
11. Ibrahim A, Refaee T, Leijenaar RTH, et al. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *PLoS One.* 2021;16(5):e0251147. doi:10.1371/journal.pone.0251147
12. Lu L, Ehmke RC, Schwartz LH, Zhao B. Assessing Agreement between Radiomic Features Computed for Multiple CT Imaging Settings. *PLoS One.* 2016;11(12):e0166550. doi:10.1371/journal.pone.0166550
13. Zhao B, Tan Y, Tsai W-Y, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep.* 2016;6:23428. doi:10.1038/srep23428
14. Fortin J-P, Parker D, Tuñç B, et al. Harmonization of multi-site diffusion tensor imaging data. *Neuroimage.* 2017;161:149-170. doi:10.1016/j.neuroimage.2017.08.047
15. Fortin J-P, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage.* 2018;167:104-120. doi:10.1016/j.neuroimage.2017.11.024
16. Da-Ano R, Masson I, Lucia F, et al. Performance comparison of modified ComBat for harmonization

- of radiomic features for multicenter studies. *Sci Rep.* 2020;10(1):10248. doi:10.1038/s41598-020-66110-w
17. Mali SA, Ibrahim A, Woodruff HC, et al. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *J Pers Med.* 2021;11(9). doi:10.3390/jpm11090842
 18. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007;8(1):118-127. doi:10.1093/biostatistics/kxj037
 19. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: Bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol.* 2019;19:33-38. doi:10.1016/j.ctro.2019.07.003
 20. Larue RTHM, van Timmeren JE, de Jong EEC, et al. Influence of gray level discretization on radiomic feature stability for different CT scanners, tube currents and slice thicknesses: a comprehensive phantom study. *Acta Oncol.* 2017;56(11):1544-1553. doi:10.1080/0284186X.2017.1351624
 21. Mackin D, Fave X, Zhang L, et al. Credence Cartridge Radiomics Phantom CT Scans - The Cancer Imaging Archive (TCIA) Public Access - Cancer Imaging Archive Wiki. *Cancer Imaging Archive.* Published online 2017. doi:10.7937/K9/TCIA.2017.zuzrml5b
 22. Clark K, Vendt B, Smith K, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging.* 2013;26(6):1045-1057. doi:10.1007/s10278-013-9622-7
 23. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Res.* 2017;77(21):e104-e107. doi:10.1158/0008-5472.CAN-17-0339
 24. Team RC. R language definition. *Vienna, Austria: R foundation for statistical computing.* Published online 2000. <http://mirror.costar.sfu.ca/mirror/CRAN/doc/manuals/R-lang.pdf>
 25. Gandrud C. *Reproducible Research with R and R Studio.* CRC Press; 2013. <https://play.google.com/store/books/details?id=u-nuzKGvoZwC>
 26. Meijering EHW, Niessen WJ, Pluim JPW, Viergever MA. Quantitative Comparison of Sinc-Approximating Kernels for Medical Image Interpolation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI'99.* Springer Berlin Heidelberg; 1999:210-217. doi:10.1007/10704282_23
 27. Lin LI. A concordance correlation coefficient to evaluate reproducibility. *Biometrics.* 1989;45(1):255-268. <https://www.ncbi.nlm.nih.gov/pubmed/2720055>
 28. Stevenson M, Stevenson MM, BiasedUrn I. Package “epiR.” Published online 2020. <ftp://ftp.sam.math.ethz.ch/sfs/pub/Software/CRAN/web/packages/epiR/epiR.pdf>
 29. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci.* 2003;43(6):1947-1958. doi:10.1021/ci034160g
 30. Zar JH. Spearman Rank Correlation. In: *Encyclopedia of Biostatistics.* John Wiley & Sons, Ltd; 2005. doi:10.1002/0470011815.b2a15150
 31. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning

- algorithms. *Pattern Recognit.* 1997;30(7):1145-1159. doi:10.1016/S0031-3203(96)00142-2
32. Parikh R, Mathai A, Parikh S, Chandra Sekhar G, Thomas R. Understanding and using sensitivity, specificity and predictive values. *Indian J Ophthalmol.* 2008;56(1):45-50. doi:10.4103/0301-4738.37595
33. Ibrahim A, Widaatalla Y, Refaee T, et al. Reproducibility of CT-Based Hepatocellular Carcinoma Radiomic Features across Different Contrast Imaging Phases: A Proof of Concept on SORAMIC Trial Data. *Cancers* . 2021;13(18). doi:10.3390/cancers13184638



PART IV

A large, abstract blue watercolor splash with a white number 12 overlaid in the center. The splash is composed of various shades of blue, from light to dark, with a textured, organic appearance. The number 12 is rendered in a clean, white, sans-serif font, positioned centrally within the splash.

12

Chapter 12

Deep learning-based classification of metastatic foci on bone scintigraphy

Authors

Abdalla Ibrahim, Akshayaa Vaidyanathan, Sergey Primakov, Flore Belmans,
Fabio Bottari, Turkey Refaee, Pierre Lovinfosse, Alexandre Jadoul,
Celine Derwael, Fabian Hertel, Henry C. Woodruff, Helle D. Zacho, Sean Walsh,
Wim Vos, Mariaelena Occhipinti, François-Xavier Hanin, Philippe Lambin,
Felix M. Mottaghy and Roland Hustinx

Adapted from

Artificial intelligence in Medicine. Submitted.

Abstract

Background and Objectives: Metastatic bone disease (MBD) is the most common form of metastases, most frequently deriving from prostate cancer. MBD is screened with bone scintigraphy (BS), which have high sensitivity but low specificity for the diagnosis of MBD, often requiring further investigations. Deep learning (DL) - a machine learning technique designed to mimic human neuronal interactions- has shown promise in the field of medical imaging analysis for different purposes, including segmentation and classification of lesions. In this study, we aim to develop a DL algorithm that can classify areas of increased uptake on bone scintigraphy scans, with automated reporting of the body region containing the lesion(s).

Methods: We collected 2365 BS from three European medical centres. The model was trained and validated on 1203 and 164 BS scans respectively. Furthermore we evaluated its performance on an external testing set composed of 998 BS scans. We further aimed to enhance the explainability of our developed algorithm, using activation maps. We compared the performance of our algorithm to that of 6 nuclear medicine physicians.

Results: The developed DL based algorithm is able to detect MBD on BSs, with high specificity and sensitivity (0.80 and 0.82 respectively on the external test set), in a shorter time compared to the nuclear medicine physicians (2.5 minutes for AI and 30 minutes for nuclear medicine physicians to classify 134 BSs), that could be applied to any BS regardless of the patient's gender and history of cancer. Further prospective validation is required before the algorithm can be used in the clinic.

Keywords:

Deep learning, Metastatic Bone Disease, Bone scintigraphy, Activation maps

Introduction

Metastatic bone disease (MBD) is the most common form of metastatic lesions (1,2). The incidence of bone metastasis varies depending on the cancer type (3), yet around 80% of MBD arise from breast and prostate cancers (4). MBD, as the name implies, is due to the propensity of these tumours to metastasize to bones, and it results in eventually difficulty treating painful lesions. Henceforth, early diagnosis is necessary for individualized management that could significantly improve a patient's quality of life (5).

MBD is usually detected using radionuclide bone scintigraphy (or bone scans, BS). BS are nuclear medicine images, which are used frequently to evaluate the distribution of active bone formation, related to benign or malignant processes, in addition to physiological processes. BS scans are indicated in a spectrum of clinical scenarios including exploring unexplained symptoms, diagnosing a specific bone disease or trauma, and the metabolic assessment of patients prior to and during the treatment(6,7). BS combining whole-body planar images and tomographic acquisition (SPECT – single photon emission computed tomography) on selected body parts are highly sensitive, as they detect metabolic changes earlier than conventional radiologic images, with lower sensitivity to lytic lesions. However, depending on the pattern it may lack the specificity to identify the underlying causes. Therefore, a SPECT/CT that correlates the findings of bone scintigraphy anatomically is often useful and leads to a more specific diagnosis of the changes noted (8), although MRI scans may also be additionally requested to clarify the diagnosis. Hence, a tool to improve the specificity of decisions based on BS, and reduce the need for further imaging is a relevant unmet clinical need.

Deep learning (DL) is a branch of machine learning (ML), and refers to data driven modelling techniques, which applies the principles of simplified neuron interactions (9). The application of imaging analysis techniques using artificial neurons on medical imaging started to draw attention decades ago (10), but it only became a major research focus recently due to the advancement in computational capacities and imaging techniques (11,12). The artificial neuron model is used as a foundation unit to create complex chains of interactions - DL layers. These layers are used to generate even more complex structures - DL architectures. The neural network (NN) training procedure is typically a cost-function minimization process. The cost function measures the error of predictions based on the ground truth labels (13), and the DL network learns how to solve a problem directly from existing data, and apply it to data it has never seen. These complex models contain the parameters (weights) for millions of neurons, which can be trained for the recognition of problem-related patterns in the data being analysed.

Several studies investigated the potential of DL-based algorithms for analysing bone

scintigraphy scans (14–16). The majority of these studies applied DL-algorithms on BS scans of diagnosed (specific) cancer patients, which could limit the learning ability of the DL-algorithm to differentiate MBD from other bone diseases. To the best of our knowledge, no study combined both male and female patients, with no-cancer patients included.

In this study, we hypothesize that DL-based algorithms can learn the pattern of metastatic bone disease on bone scintigraphy scans, and differentiate it from other non-metastatic bone diseases. We investigate the potential of a DL-based algorithm to detect MBD on BS not limited to those of cancer patients using weakly-supervised detection based on activation maps obtained using the gradient weighted class activation mapping (Grad-CAM) method (17,18). By doing so, we aim to develop a generalizable tool that can classify scans containing metastases and detect MBD on BS, regardless of the gender and malignancy status of the patient. Moreover, extracting activation maps with the Grad-CAM method (19) and superimposing these maps to the original BD scans, we explored the explainability of the deep learning model's predictions. This is very important to promote the application of these methods in the clinic and avoid the common misconception that sees DL models as “black boxes” without any real connection to clinical and imaging characteristics. As a complementary step, we explored the development of an automated label generator for the location of the detected metastatic foci.

Materials and Methods

Imaging data

The imaging data were retrospectively collected from different European centres: Aachen RWTH University Clinic (Aachen, Germany), Aalborg University Hospital (Aalborg, Denmark), and Namur University Hospital (Namur, Belgium). The electronic medical records of these hospitals were searched for patients who underwent BS between 2010 and 2018. Patients for whom a definitive classification of the foci was available, mostly through further investigations, were further included. All images were acquired with anteroposterior (AP) and posteroanterior (PA) whole-body views. The imaging analysis was approved by the Aachen RWTH institutional review board (No. EK 260/19), and informed consent was obtained from all included patients. According to Danish National Legislation, the Danish Patient Safety Authority can waive informed consent for retrospective studies (approval 31-1521-110). All methods were carried out in accordance with the relevant guidelines and regulations (20). The study protocol was published on clinicaltrials.gov (NCT: NCT05110430)

Image pre-processing

Every datapoint containing acquisition at two views (AP and PA) was resized to size (length = 256, height = 512) and the intensities were normalized to range [0-1] using the minimum and maximum intensity of each image. For all the data points, image acquisitions at both views are appended besides each other as shown in Figure 1.

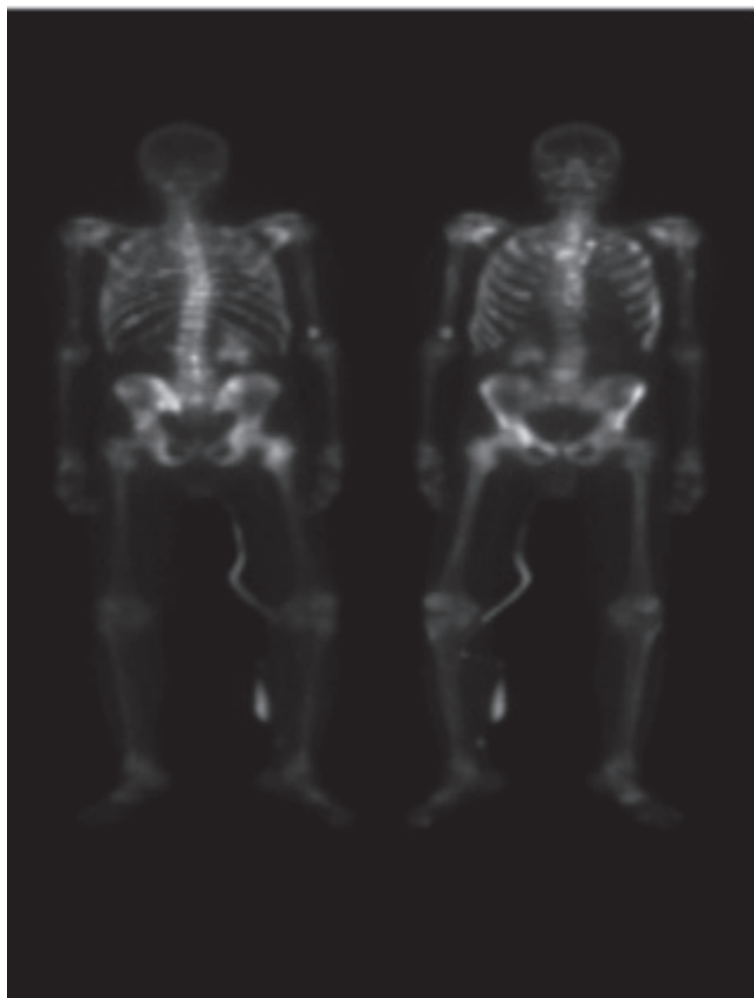


Figure 1. Example of pre-processed BS scans used as input for model training

Model architecture, training and testing

The training and validation datasets are composed of 1203 and 164 images respectively, coming from Centre A (Aachen) and B (Aalborg). The external test cohort is composed of 998 images collected at centre C (Namur). A full overview of the patients cohort division between the different datasets is reported in Table 1.

Table 1. Division of the patients cohort between training, validation and external test

	Training (n = 1203)	Validation (n = 164)	External test (n = 998)
Centre A (Achen)	235 with metastasis 668 normal	58 with metastasis 58 normal	-
Centre B (Albourg)	94 with metastasis 206 normal	24 with metastasis 24 normal	-
Centre C (Namur)	-	-	411 with metastasis 587 normal

The model was trained on 329 images containing metastasis from Centre B (94) and A (235). At each epoch, the 874 images without any metastasis were shuffled and 329 images were randomly selected to train the model with balanced labels. VGG16 architecture with ImageNet pretrained weights (21) was trained with categorical cross entropy loss for 6 epochs with 200 steps per epoch. The model was trained with 3 channel input. The pre-processed input was duplicated in all the channels. During the training, the images were augmented (22) by flipping along the vertical axis so that the views at AP and PA were randomly represented in the left or right in the images.

The last Max Pooling layer in the VGG16 model was followed by a Global Average pooling layer, followed by a fully connected layer with 512 units and ReLu activation, which is followed by a classification layer containing 2 units with Softmax activation (23) as shown in Figure 2. The network weights are updated by using the Adam optimizer at an initial learning rate of $1e^{-4}$ (24).

The trained model's performance was evaluated on an external test dataset (n = 998).

Deep learning-based classification of metastatic foci on bone scintigraphy

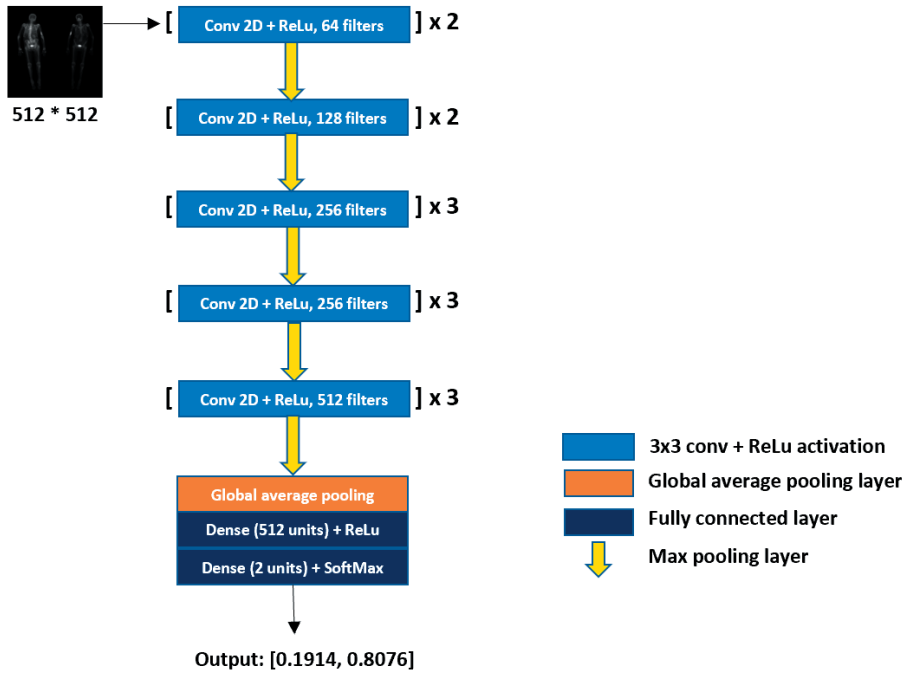


Figure 2. The architecture used in the study. Pre-processed BS scans resized to 512 * 512 dimensions were provided as input to the network. The network outputs a probability score for presence and absence of metastasis on BS images. X = block repetitions, Conv = Convolution kernel, ReLU = rectified linear unit, 3x3 = the size of the 2D CNN kernels.

Automatic labeller for the location of metastasis in bone scintigraphy scans

A dataset of BS was provided by University of Aachen and contained the scans of 20 patients, each containing both AP and PA views. All scans had annotations for six anatomical regions (head, thorax, pelvis, shoulders, upper limbs and lower limbs), as shown in Figure S1. The total of 40 scans was split into a training (32) and validation (8) set.

A ResNext50 architecture (25) with ImageNet pretrained weights (26) was trained with categorical cross-entropy loss. A 3 channel input was used where the first channel contained the scan while the two others contained a segmented region. The segmented region was artificially created from the region annotations that came with the dataset. An example of a scan with a segmented region is shown in Figure S2. The last convolutional layer in the ResNext50 model was followed by a Global Average pooling layer which reduces the image spatial resolution, followed by a fully connected layer with 512 units and ReLU activation, which is followed by a classification layer containing 6 units with Softmax activation. The network weights are updated by using the Adam optimizer at an initial learning rate of $1e^{-5}$. Due to the limited number of scans, the fact that

metastasis can occur in a lot of different locations and the fact that metastatic regions are much smaller than the region annotations, extensive augmentation was applied during training. Three different augmentations were applied during training:

1. Variation in the highlighted region of the scan (head, thorax, pelvis, shoulders, upper limbs and lower limbs)
2. Variation in the number of pixels highlighted
3. Variation in the shape of the highlighted region
4. Left/right flip of the scan

Quantitative metrics

The quantitative model performance in this study was assessed using ROC AUC, sensitivity and specificity of the classifier and confusion matrix (true positive rate (TPR), true negative rate (TNR), false negative rate (FNR) and false positive rate (FPR)). The model was evaluated according to the Checklist for AI in Medical Imaging (CLAIM) (27) and Standards for Reporting Diagnostic accuracy studies (STARD) (28).

In silico clinical trial

To better gauge the proposed DL model performance, we developed an application allowing the creation of a reference performance point by collecting nuclear medicine physician's feedback based on the visual assessment of BS scans. We have enrolled 6 nuclear medicine physicians to measure their performance on the evaluation dataset of 134 BS images. This dataset was sampled from the Centre C images with an equal number of negative and positive cases. In order to collect participant's feedback, the application was displaying BS image, comment window and window filtering settings (Fig. 3). In the end of the feedback assessment excel file was generated. For better visual comparison we have evaluated DL based AUC on the same dataset that has been used for visual assessment (134 BS images). We used bootstrapping with 100 iterations to generate DL based AUC distribution.

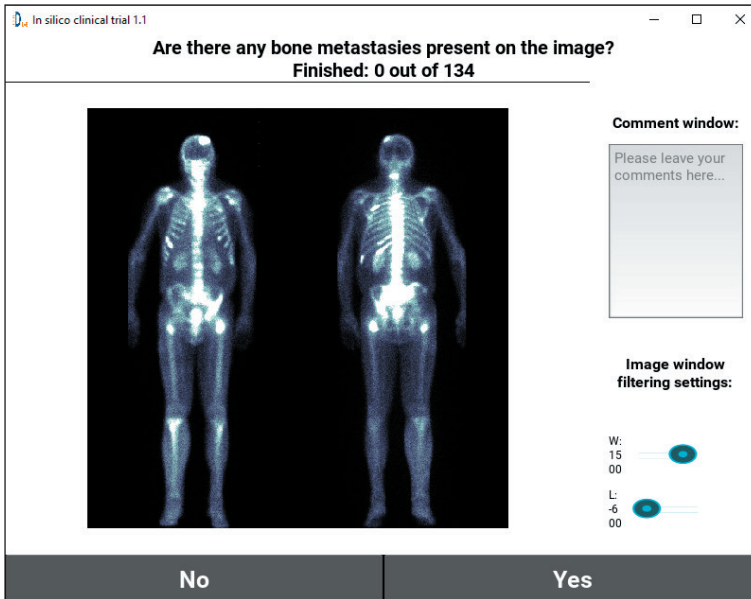


Figure 3. Screenshot of the application feedback window.

Results

Model performance

The classification performances of the DL model were evaluated on the external test set coming from Centre C, in terms of Area under the Curve (AUC). The AUC gives the diagnostic ability of a binary classifier to discriminate between true and false values, in this case metastatic and non-metastatic bone disease. Fig. 4 (left) represents the ROC curve of the DL classification model, while Fig. 4 (right) is the confusion matrix, which reports the percentages of correct and incorrect classification for each class (metastatic and non-metastatic).

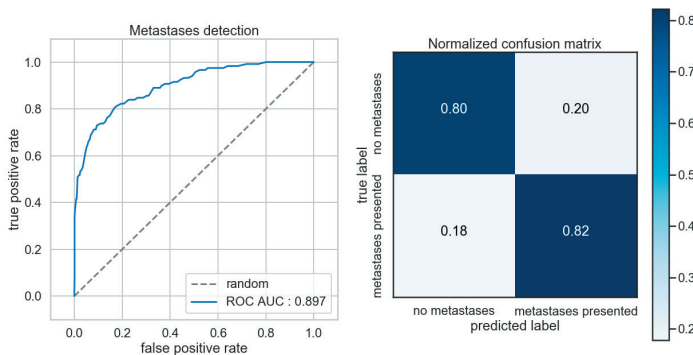


Figure 4. ROC curve for the classification DL model (left) and Confusion matrix (right)

The model achieved an AUC of 0.897, TPR of 82.2%, TNR of 80.45 %, FPR of 19.55% and FNR of 17.79 % on the external test set. The model achieved a CLAIM score of 64 % (27 out of 42 items) and STARD of 50 % (15 out of 30 items).

Explainability of trained model based on activation maps

During the testing phase of the trained model, for the scans that were predicted positive (i.e. metastatic disease), activation maps were extracted using the Grad-CAM method. The method uses the gradients extracted corresponding to the class with highest predicted probability, flowing through the last convolutional layer, to produce the activation map. The map was then resized to the size of the input image and superimposed on the original BS scan, allowing visual inspection of activated zones on the image as shown in Figure 5 and 6. The activated regions are compared with radiologist's' segmentation of metastatic spots for qualitative assessment of the explainability of the model's predictions.

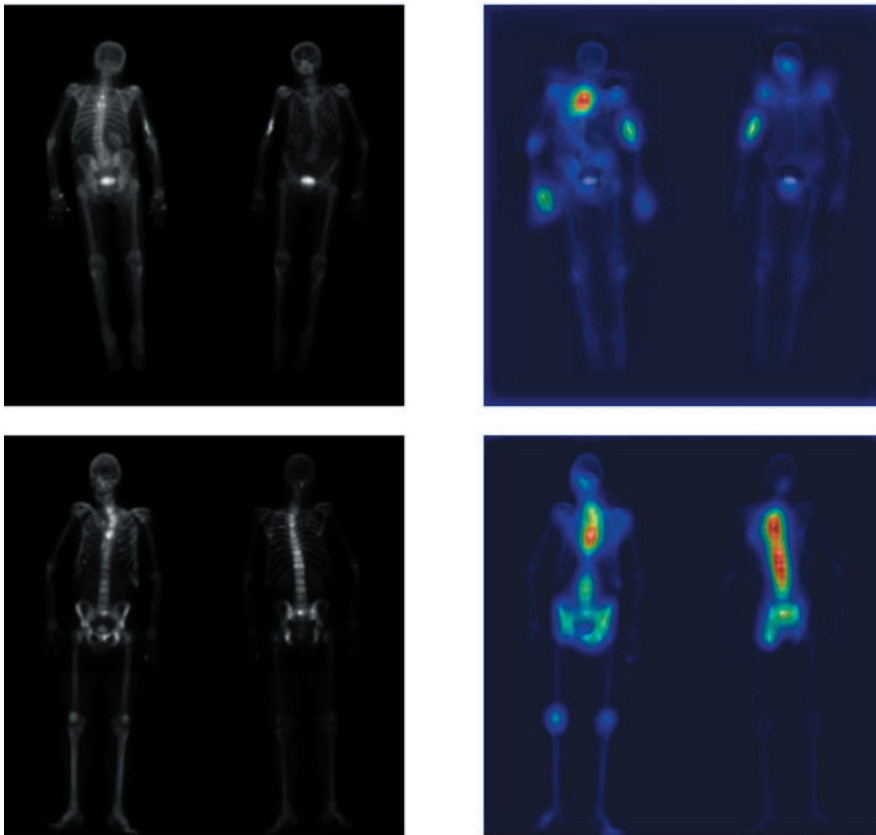


Figure 5. BS images which are correctly classified along with their corresponding activation maps extracted using the GRAD-CAM method. Left) original BD scan, Right) Grad-CAM activation maps obtained from the DL model. Scan correctly classified with a probability of 0.78 (top) and 0.99 (bottom)

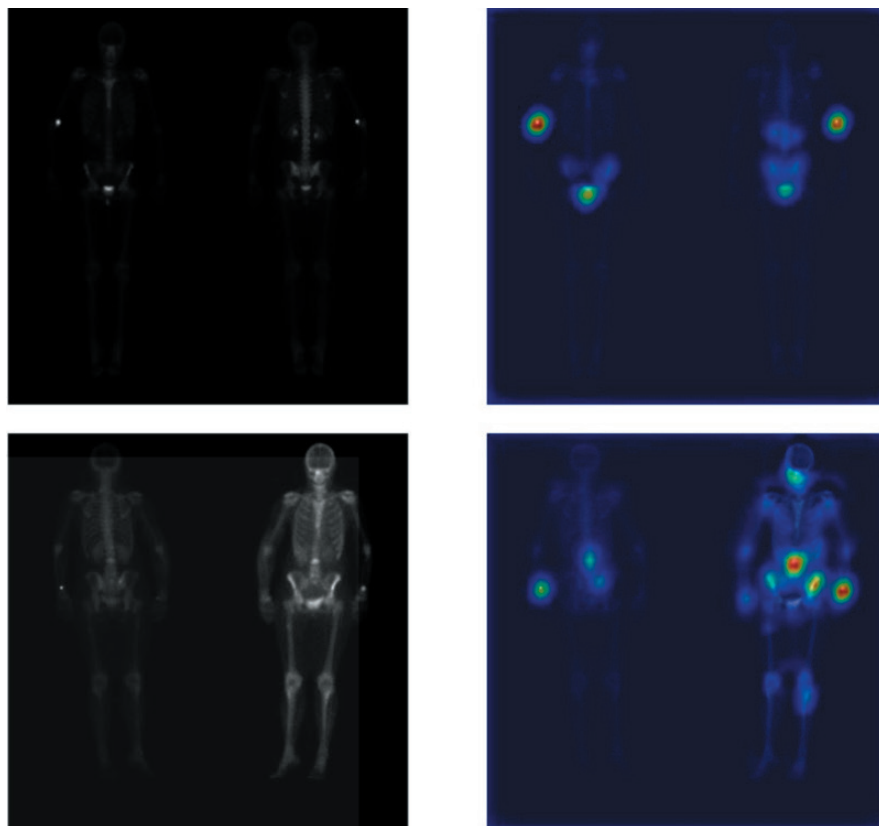


Figure. 6 BS images which are wrongly classified along with their corresponding activation maps extracted using the GRAD-CAM method. Left) original BD scan, Right) Grad-CAM activation maps obtained from the DL model. Scan incorrectly classified with a probability of 0.79 (top) and 0.63 (bottom)

Automatic labeller for the location of metastasis in bone scintigraphy scans

We developed an automatic labeller for the location of metastasis in BS, after the metastatic regions have been extracted. This objective is of great interest as it would allow automated completion of the clinical report with the location of metastasis. The approach proposed here automatically predicts the anatomic locations of metastasis in BS, given the scan and metastatic region as input. For this purpose, a model was built to distinguish between 6 different anatomic regions: head, thorax, pelvis, shoulders, upper limbs and lower limbs. At the end of training, a categorical accuracy of 0.92 was reached on the validation set. However, segmented spots for the scans in the validation set were also artificially created. The trained model was therefore tested on an external dataset ($n = 462$) of BS scans with indications of metastatic regions extracted from the activation maps of the MBD classifier. The resulting labels were qualitatively evaluated. A few examples are shown in Figures S3 and S4.

In silico clinical trial

The performance of nuclear medicine physicians based on the BS images was evaluated using AUC, where median performance of the nuclear medicine physician was 0.895 (IQR = 0.087) and median performance of DL based method was 0.95 (IQR = 0.024) (Fig. 7).

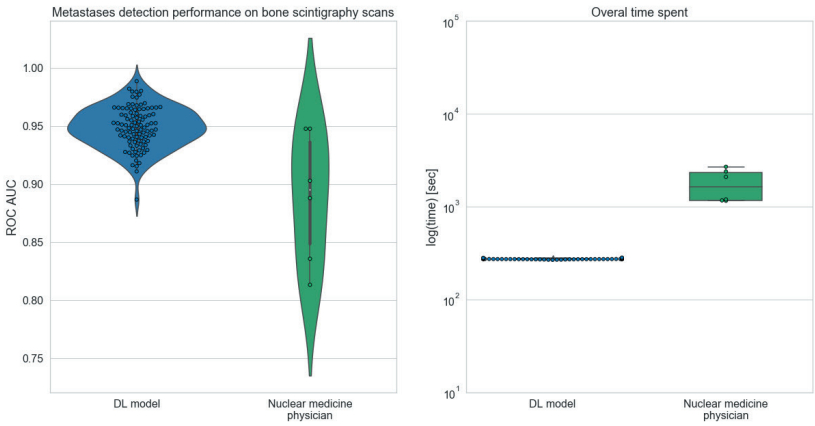


Figure. 7 Violin plots showing the distributions of AUC scores for DL based and manual (across physicians) metastases detection on BS (left); boxplots of the log of the time needed by DL algorithm and nuclear medicine physicians (right).

On average, nuclear medicine physicians spent 30 mins on average to classify all the 134 scans. Given that the physicians had no access to clinical information about the patients, it takes on average 15 seconds to review one scan. In comparison, our developed algorithm takes 2 and half minutes to classify all the 134 scans, which is around 2 seconds per patient scan.

Discussion

In this study, we investigated the potential of DL-based algorithms to detect MBD on BSs collected from different centres without limiting the study population to cancer patients. Our results show that DL-based algorithms have a great potential to be applied as clinical decision aid tools, which could minimize the time needed by a nuclear physician to assess BSs, and increase the diagnostic specificity of BSs. The application of the state-of-the-art classification techniques has yielded a performance similar to nuclear physicians with no background about the patients’ history, which was further endorsed by the results of the in silico clinical trial.

Besides classification and the extraction of activation maps, first exploratory steps were taken towards the development of a model to automatically label for the location of metastasis which can be extended further to automatic report generation in a clinical setting. This latter objective is of great interest as it would allow to automatically complete the clinical report file with the location of metastasis. For this purpose, a classification model, based on ResNet50 architecture, was built to distinguish between 6 different anatomic regions: head, thorax, pelvis, shoulders, upper limbs and lower limbs. The ground truths to train the classifier consisted of images with indicated regions at aforementioned locations. In order to create a robust model from the available labels, augmentation techniques were applied during training. These include variation in the highlighted region of one scan, variation in the number of pixels highlighted and variation in the shape of the highlighted region (29). This preliminary work resulted in a DL model able to classify activated metastatic regions into 6 anatomical categories with performance of AUC 0.92. These preliminary results showed the potential of a DL-based classifier to automatically label the location of metastasis in bone scintigraphy scans which can be used to finalize clinical reports. However, further validation of this model is needed in the future.

Some studies previously investigated the potential of DL algorithms to classify lesions on BSs. A study investigated the potential of a DL algorithm trained on 139 patients to detect MBD on BSs of prostate cancer patients (16). The authors reported that the nuclear medicine physicians participating in the study achieved a higher sensitivity and specificity compared to the DL algorithm, though the differences were not statistically significant, and highlighted the possibility of involving DL in this clinical aspect. Another study also investigated the ability of DL algorithms to detect MBD in BS of prostate cancer patients (15). The authors trained the algorithm on 778 BS that could accurately (accuracy of $91.61\% \pm 2.46\%$) detect MBD for prostate cancer patients on BS. However, the authors did not report on the comparison with the performance of nuclear medicine physicians. Another study investigated the performance of two DL architectures for classifying BS of prostate cancer patients (30). The study included a large number of scans, and the authors reported that the best model achieved an overall accuracy of 0.9. Anand et al. reported on the performance of EXINI bone software, a classification tool for classifying BS of prostate cancer patients based on bone scan index, on simulated and patient scans (31). The authors reported that the software was more consistent in classifying BS compared to visual assessment. Uniquely, we trained our model on patients with and without a history of cancer. The use of our developed algorithm resulted in better classification results on the external test set compared to the median nuclear medicine physician performance, in a significantly shorter time. These results highlight the potential of such algorithms to become reliable clinical decision support tools that minimize the time a clinician needs to review bone scintigraphy

scans. Furthermore, our automatic labelling function and the Grad-CAM maps allow the nuclear physicians to rapidly check the spots based on which the classification was made.

While our study included a relatively large number of scans for training and externally testing the algorithm, several limitations of this study should be noted. Although explainability of model's predictions were explored with qualitative assessment, this study lacks quantitative assessment of the activations due to the limited number of manual segmentations of metastasis (c.a. 25) on the external test dataset. Also, as shown in figure 7, the activated zones correspond to the injected spot in the hand, which shows model's overfitting (32) on features that are not relevant to the metastatic spot to classify presence or absence of metastasis in images. Secondly, a prospective validation is required to properly assess the impact of using the algorithm on the current standard of care. Lastly, the physicians' performances in the in silico trial are only indicative, as they dealt with planar images only, without SPECT and CT, and without any clinical input. Obviously, this merely approximates the actual routine clinical setting, but it provides a fair indication of the potential added value of DL in this setting.

In conclusion, we developed a DL based algorithm that is able to detect MBD on BSs, with high specificity and sensitivity, that could be applied to any BS regardless of the patient's gender and history of cancer. Further prospective validation is required before the algorithm can be used in the clinic.

KEY POINTS

QUESTION: The accurate and time sensitive classification of metastatic foci on bone scintigraphy scans

PERTINENT FINDINGS: The DL mode trained on retrospective data from three different centers is able to detect MBD on BSs, with high specificity and sensitivity (80 % and 82 % respectively), in a shorter time compared to the nuclear medicine physicians

IMPLICATIONS FOR PATIENT CARE: Our algorithm might become a reliable clinical decision support tool, shortening the time a clinician needs to review BS scans and, thanks to Grad-CAM maps, rapidly review the area of the scan on which the classification was made.

Funding

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2020-PoC: 957565-AUTO.DISTINCT, Authors also acknowledge financial support from the European Union's Horizon 2020 research and

innovation programme under grant agreement: ImmunoSABR n° 733008, MSCA-ITN-PREDICT n° 766276, CHAIMELEON n° 952172, EuCanImage n° 952103, Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4) and Maastricht-Liege Imaging Valley grant, project no. “DEEP-NUCLE”.

Authors' contribution

Abdalla Ibrahim, Akshayaa Vaidyanathan: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization
Sergey Primakov, Flore Belmans, Fabio Bottari, Turkey Reface: Methodology, Validation, Formal analysis, Data Curation, Writing - Review & Editing, Visualization

Pierre Lovinfosse, Alexandre Jadoul, Celine Derwael, Fabian Hertel, Helle D. Zacho, Francois-Xavier Hanin: Validation, Investigation, Resources, Data Curation, Writing - Review & Editing

Henry C. Woodruff, Sean Walsh, Wim Vos, Mariaelena Occhipinti, FX Hanin: Resources, Writing - Review & Editing

Philippe Lambin, Felix M. Mottaghy, Roland Hustinx: Conceptualization, Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

Conflicts of interest

Akshayaa Vaidyanathan, Flore Belmans and Fabio Bottari are salaried employees of Radiomics.

Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Radiomics SA, ptTheragnostic/DNAmito, and Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/consultancy fee and/or in kind manpower contribution from Radiomics SA, BHV, Merck, Varian, Elekta, ptTheragnostic, BMS and Convert pharmaceuticals. Lambin has minority shares in the companies Radiomics SA, Convert pharmaceuticals, Comunicare, and LivingMed Biotech, and he is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics SA and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics SA and Health Innovation Ventures and three non-issues, non-licensed patents on Deep LearningRadiomics and LSRT (N2024482, N2024889, N2024889). He confirms that none of the above entities or funding was involved in the preparation of this paper.

Henry Woodruff has minority shares in the company Radiomics.

Felix M. Mottaghy received an advisor fee and reimbursement of travel costs from Radiomics. He reports institutional grants from GE and Nanomab outside the submitted work.

Mariaelena Occhipinti reports personal fees from Radiomics, outside the submitted work

Wim Vos and Sean Walsh have shares in the company Radiomics.

The rest of co-authors declare no competing interest.

References

1. Coleman RE. Clinical features of metastatic bone disease and risk of skeletal morbidity. *Clin cancer Res an Off J Am Assoc Cancer Res.* 2006;12:6243s-6249s.
2. Migliorini F, Maffulli N, Trivellas A, Eschweiler J, Tingart M, Driessen A. Bone metastases: a comprehensive review of the literature. *Mol Biol Rep.* 2020;47:6337-6345.
3. Huang J-F, Shen J, Li X, et al. Incidence of patients with bone metastases at diagnosis of solid tumors in adults: a large population-based study. *Ann Transl Med.* 2020;8:482.
4. Coleman RE. Metastatic bone disease: clinical features, pathophysiology and treatment strategies. *Cancer Treat Rev.* 2001;27:165-176.
5. Macedo F, Ladeira K, Pinho F, et al. Bone Metastases: An Overview. *Oncol Rev.* 2017;11:321.
6. Ryan PJ, Fogelman I. Bone scintigraphy in metabolic bone disease. *Semin Nucl Med.* 1997;27:291-305.
7. Ziessman HA, O'Malley JP, Thrall JHBT-NM (Fourth E, eds. Chapter 7 - Skeletal Scintigraphy. In: Philadelphia: W.B. Saunders; 2014:98-130.
8. Van den Wyngaert T, Strobel K, Kampen WU, et al. The EANM practice guidelines for bone scintigraphy. *Eur J Nucl Med Mol Imaging.* 2016;43:1723-1738.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436-444.
10. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys.* 1943;5:115-133.
11. Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Trans Signal Inf Process.* 2014;3:e2.
12. Aslam Y, N S. A Review of Deep Learning Approaches for Image Analysis. In: 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT). ; 2019:709-714.
13. Janocha K, Czarnecki WM. On loss functions for deep neural networks in classification. *Schedae Informaticae.* 2016;25:49-59.
14. Cheng D-C, Hsieh T-C, Yen K-Y, Kao C-H. Lesion-Based Bone Metastasis Detection in Chest Bone Scintigraphy Images of Prostate Cancer Patients Using Pre-Train, Negative Mining, and Deep Learning. *Diagnostics.* 2021;11.
15. Papandrianos N, Papageorgiou E, Anagnostis A, Papageorgiou K. Efficient Bone Metastasis Diagnosis in Bone Scintigraphy Using a Fast Convolutional Neural Network Architecture. *Diagnostics (Basel, Switzerland).* 2020;10.
16. Aoki Y, Nakayama M, Nomura K, et al. The utility of a deep learning-based algorithm for bone scintigraphy in patient with prostate cancer. *Ann Nucl Med.* 2020;34:926-931.
17. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Int J Comput Vis.* 2020;128:336-359.
18. Dubost F, Adams H, Yilmaz P, et al. Weakly supervised object detection with 2D and 3D regression neural networks. *Med Image Anal.* 2020;65:101767.
19. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *Int J Comput Vis.* 2016;128:336-359.

20. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*. 2013;310:2191-2194.
21. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *3rd Int Conf Learn Represent ICLR 2015 - ConfTrack Proc*. 2015:1-14.
22. Shorten C, Khoshgoftaar TM. A survey on Image Data Augmentation for Deep Learning. *J Big Data*. 2019;6:60.
23. Calin O. Activation Functions BT - Deep Learning Architectures: A Mathematical Approach. In: Calin O, ed. Cham: Springer International Publishing; 2020:21-39.
24. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. ; 2015.
25. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. Vol 2017-Janua. Institute of Electrical and Electronics Engineers Inc.; 2017:5987-5995.
26. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. ; 2009:248-255.
27. Mongan J, Moy L, Kahn CE. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers. *Radiol Artif Intell*. 2020;2:e200029.
28. Cohen JF, Korevaar DA, Altman DG, et al. STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. 2016;6:e012799.
29. Abdollahi B, Tomita N, Hassanpour S. Data Augmentation in Training Deep Learning Models for Medical Image Analysis BT - Deep Learners and Deep Learner Descriptors for Medical Applications. In: Nanni L, Brahnam S, Brattin R, Ghidoni S, Jain LC, eds. Cham: Springer International Publishing; 2020:167-180.
30. Han S, Oh JS, Lee JJ. Diagnostic performance of deep learning models for detecting bone metastasis on whole-body bone scan in prostate cancer. *Eur J Nucl Med Mol Imaging*. August 2021.
31. Anand A, Morris MJ, Kaboteh R, et al. Analytic Validation of the Automated Bone Scan Index as an Imaging Biomarker to Standardize Quantitative Changes in Bone Scans of Patients with Metastatic Prostate Cancer. *J Nucl Med*. 2016;57:41-45.
32. M.R.Narasinga Rao D, Venkatesh Prasad V, Sai Teja P, Zindavali M, Phanindra Reddy O. A Survey on Prevention of Overfitting in Convolution Neural Networks Using Machine Learning Techniques. *Int J Eng Technol*. 2018;7:177.



13

Chapter 13

Validated fully automated detection
and segmentation of non-small cell lung
cancer on computed tomography images

Authors

Sergey P. Primakov, Abdalla Ibrahim, Janita E. van Timmeren,
Guangyao Wua, Simon A. Keek, Manon Beuque, R. Granzier, Madeleine Scrivener,
Sebastian Sanduleanu, Esmā Kayan, Iva Halilaj, Jianlin Wu, René Monshouwer,
Hester A Gietema, Lizza E.L. Hendriks, Olivier Morink, Arthur Jochemsa,
Henry C. Woodruff and Philippe Lambina

Adapted from

Nature Communications. Accepted.

Abstract

Detection and segmentation of abnormalities on medical images is highly important for patient management including diagnosis, radiotherapy, response evaluation, as well as for quantitative image research.

We developed and validated a fully automated pipeline for the detection and volumetric segmentation of non-small cell lung cancer (NSCLC) using 1343 thoracic CT scans from 8 institutions. Along with quantitative performance detailed by image slice thickness, tumor size, and image interpretation difficulty, we have performed an “*in silico*” prospective clinical trial, which showed that the proposed method was faster and more reproducible compared to the experts. On average, radiologists & radiation oncologists preferred automatic segmentations in 56% of the cases.

Additionally, we evaluated the prognostic power of the automatic contours by applying RECIST criteria and measuring the tumor volumes. Segmentations by our method stratified patients into low and high survival groups with higher significance compared to those methods based on manual contours.

Introduction

Lung cancer is the deadliest of all cancers afflicting both sexes, accounting for 18.4% of the total cancer deaths worldwide in 2018, almost equal to breast and colon cancers combined ¹. Recent advances in treatment (immune checkpoint inhibitors, tyrosine kinase inhibitors) has significantly improved survival times for subgroups of patients. However, much work is still to be done in the field of lung cancer, especially in screening and early detection. Automated detection and segmentation would immediately impact the clinical workflow in radiotherapy, one of the most common treatment modalities for lung cancer ². Radiotherapy uses medical imaging, especially computed tomography (CT), to obtain accurate tumor localization and electron densities for the purpose of treatment planning dose calculations ³. Accurate segmentations of the tumor and organs at risk are also essential as errors might lead to over- or under-irradiation of both the tumor and/or healthy tissue. It has been estimated that a 1mm shift of the tumor segmentation could affect the radiotherapeutic dose calculations by up to 15% ^{4,5}.

Equally important are the lesion and organ at risk segmentation process for radiation oncologists for radiotherapy planning, and the measurement of lesions within the Response Evaluation Criteria in Solid Tumors (RECIST) 1.1 framework for radiologists, both laborious manual routines which impose an avoidable workload ⁶. Currently, such segmentations and appropriate RECIST measurements are performed manually or semi-automatically, consuming valuable time and resources, as well as being prone to inter- and intra-observer variability ⁷.

Another field to profit directly from automated detection and delineation of lesions is radiomics, the high-throughput mining of quantitative features from medical images and their subsequent correlation with clinical and/or biological endpoints ^{8,9}. Radiomics has the potential to facilitate personalized medicine via diagnostic and predictive models based on phenotypic properties of the region of interest (ROI) being analyzed ¹⁰. ROI segmentation is currently considered to be one of the most time intensive and laborious steps within the entire radiomics workflow ¹¹.

Taking into consideration these clinical and research needs for lung tumor segmentation, the implementation of automated detection software that is capable of fast and accurate delineation of NSCLC on thoracic CT scans is desirable, bordering on necessary. The applications and benefits include, but are not limited to: 1) CT-based automated screening of lung cancer; 2) Retrospective analysis of entire databases of patients who underwent thoracic CT in daily care for research purposes; 3) Consistent and reproducible segmentations, which are important in planning and monitoring (radio)therapy, and in research; 4) Follow-up of treated primary lung cancer; 5) Automation and acceleration

of certain aspects of the clinical radiotherapy workflow, making adaptive re-planning more feasible.

The recent advancement of machine learning techniques, combined with improvements in the quality and archiving of medical images, have fueled intensive research in the field of artificial intelligence (AI) for medical imaging analysis^{12,13}. Deep learning, a branch of AI based on artificial neural networks, has been successfully applied on images to solve problems such as classification or segmentation^{14,15}. Several attempts have been made to adapt these methods for medical imaging problems, including tumor detection and segmentation on CT images¹⁶⁻¹⁹. A major hurdle in developing fully automated software that can be applied to any CT is the heterogeneity of the datasets, especially when acquired from multiple centers²⁰. CT scans with different acquisition- or reconstruction parameters present lung structures differently. The methods described in the current literature usually lack a CT pre-processing module in the pipeline, and the problem of data harmonization is left to be solved by a data driven approach, requiring large datasets representing all aspects of this inhomogeneity.

The aim of this study was to develop a fully automated lung tumor detection and 3D volumetric segmentation pipeline that is capable of handling a large variety of CT acquisition and reconstruction parameters. Furthermore, our model was validated on 3 external datasets and a volumetric prognostic factor was compared to an existing clinical standard and to a similar published method. We have also performed an “*in silico*” prospective clinical trial to compare speed and reproducibility of our method to those of experts.

Results

Overall, 1328 thoracic volumetric CT scans with corresponding 3-dimensional tumor segmentations were used in order to train, test and validate a fully automated method for detection and segmentation of NSCLC in standard-of-care images. Datasets 1-7 were combined and randomly divided into training and testing datasets with 999 patients and 93 patients, respectively (see Table 1). Datasets 8-10, comprising 236 patients were used for external validation of the method. A summary of the data is provided in Table 1.

Tumor detection and segmentation

A 3-step workflow was developed and successfully implemented (Fig.1): (i) image pre-processing, a crucial step as datasets collected for this work were obtained from different scanners with various image acquisition and reconstruction protocols (Figure 1 suppl.). The data inhomogeneity necessitated the harmonization of CT data in order

to achieve comparable representations of the tumor region, reduce computational power requirements and image noise, and to optimize contrast; (ii) lung isolation, which allows the model to focus on the ROI and the input of the entire CT scans; (iii) automated tumor detection and segmentation, employing the convolutional neural network.

The ability of the system to detect tumors was assessed slice-wise and yielded a median accuracy of 0.93 in the validation dataset, and a median area under the receiver operating characteristic curve (AUC) of 0.89. The median contouring performance in the validation dataset as assessed by the volumetric Dice similarity coefficient (DSC) was 0.77, while the Jaccard index (JI) was 0.62. Further metrics, associated uncertainties, as well as test dataset results are reported in Table 2.

Model performance was also separately assessed in regard to groupings of image slice-thickness, tumor size, and expert-reported tumor complexity. The sub-cohorts were analyzed for significant differences in model performance, with the results reported in Table 3. As some of the tumors had two or more unconnected components (Satellite lesions, or edges of the tumor), the Hausdorff metric can yield unreliable distances when the distance between different volume fragments are calculated. Therefore, the IQR for H95th was not provided. Histograms showing the distributions of detection and segmentation results are provided in the supplementary materials (Fig.2 suppl. and Fig.3 suppl.).

Boxplots showing DSC distributions in the sub cohorts tumor size and tumor complexity for both test and validation datasets are shown in Figure 2. There is a clear trend toward better performance and less variability for larger and less complex tumors. More comparisons for differing slice-thickness groups, complexity classes, and tumor sizes performed on the validation dataset are provided in the supplementary materials (Fig. 4-6 suppl.).

Examples of the automatically generated segmentations (from the validation set) in comparison to contours segmented by experts are shown in Figure 3.

Table 1. Description of the datasets used in this study

Ref. #	#Dataset	Dataset	Use	Medical Center	#Patients used	#CT slices with tumor (%)	#CT slices without tumor (%)*	Mean Tumor volume (ml)
21	1	Maastrro-CT-Lung-1	Training/Test	Open source (TCIA)	422	4262 (16)	22490 (84)	71.0
N/A	2	UCL-CT-Lung	Training/Test	Université catholique de Louvain	39	400 (16)	2096 (84)	53.44
N/A	3	UCSF-CT-Lung	Training/Test/ Clinical trial	University of California - San Francisco	101	689 (11)	5775 (89)	19.35
N/A	4	MUMC+ Inoperable Lung	Training/Test	Maastricht University Medical Center+	92	1247 (21)	4577 (79)	94.99
N/A	5	AZHDU Lung	Training/Test	Affiliated Zhongshan Hospital of Dalian University	222	464 (4)	9456 (96)	2.08
22	6	Stanford Lung	Training/Test	Open source	137	796 (10)	7396 (90)	22.37
23	7	TCIA-CT-Lung-3	Training/Test	(TCIA)	92	630 (12)	4618 (88)	51.39
24	8	The Maastrro interobserver reproducibility test	Validation	Open source	22	210 (16)	1070 (84)	88.03
N/A	9	Radbound Lung 2	Validation	(BMIA XNAT) Radbound University Medical Center	132	3493 (22)	12460 (78)	92.04
N/A	10	MUMC/Heerlen Lung	Validation	MUMC/Heerlen	84	1120 (13)	7317 (87)	77.79
		Overall training/test			1105	8488 (13)	56408 (87)	49.07
		Overall validation			236	4823 (19)	20843 (81)	88.93

*CT slices without a segmentation were considered as not containing tumor

Validated fully automated detection and segmentation of non-small cell lung cancer

Table 2. Overview of quantitative model performance. IQR = Interquartile range, DSC = Dice similarity coefficient, Ji = Jaccard index, H95th = 95th percentile Hausdorff distance.

Data, # of patients	Detection performance				Segmentation performance		
	Accuracy (IQR)	Slicewise AUC (CI)	Specificity	Sensitivity	DSC (IQR)	Ji (IQR)	H95th, mm
Test, 93	0.94 (0.08)	0.89 (0.89- 0.90)	0.89	0.90	0.77 (0.24)	0.63 (0.32)	6
Validation, 236	0.93 (0.08)	0.89 (0.89- 0.90)	0.92	0.86	0.77 (0.23)	0.62 (0.29)	10

Table 3. Overview of quantitative model performance with regard to various factors

Factors	Test				Validation				
		DSC (IQR)	Significance		DSC (IQR)	Significance			
Slice thickness, [mm]	0-2.5	0.76 (0.24)	-	ns	ns	0.75 (0.23)	-	ns	ns
	2.5-5	0.76 (0.28)	ns	-	-	0.76 (0.21)	*	-	-
	>5	0.80 (0.19)	-	ns	ns	0.82 (0.20)	-	-	ns
Complexity label (need PET)	0	0.83 (0.21)	****	-	-	0.81 (0.15)	****	-	-
	1	0.76 (0.26)	-	-	-	0.76 (0.22)	-	-	-
	<20	0.73 (0.28)	-	ns	ns	0.72 (0.23)	-	*	****
Tumor size, [ml]	20-150	0.79 (0.20)	ns	-	-	0.76 (0.21)	**	-	-
	>150	0.88 (0.12)	-	ns	ns	0.83 (0.15)	-	-	****

ns (non significant): $5.00e-02 < p \leq 1.00e+00$ / *: $1.00e-02 < p \leq 5.00e-02$ / **: $1.00e-03 < p \leq 1.00e-02$ / ***: $1.00e-04 < p \leq 1.00e-03$ / ****: $p \leq 1.00e-04$

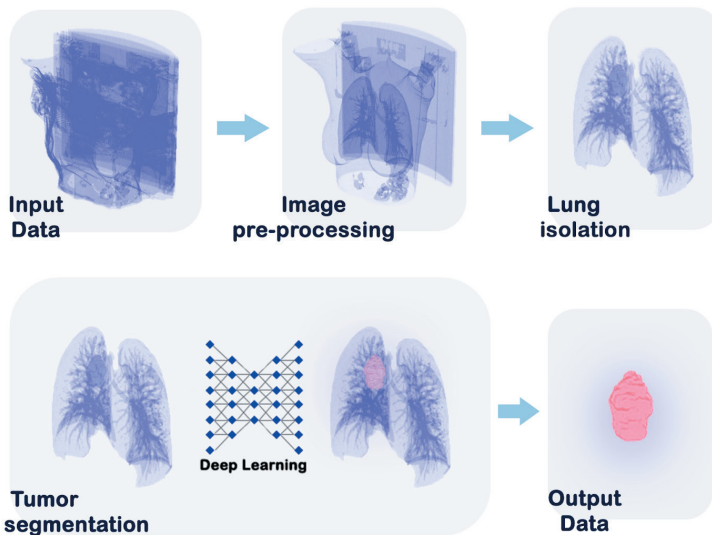


Figure 1. Graphic representation of the major steps in the proposed workflow

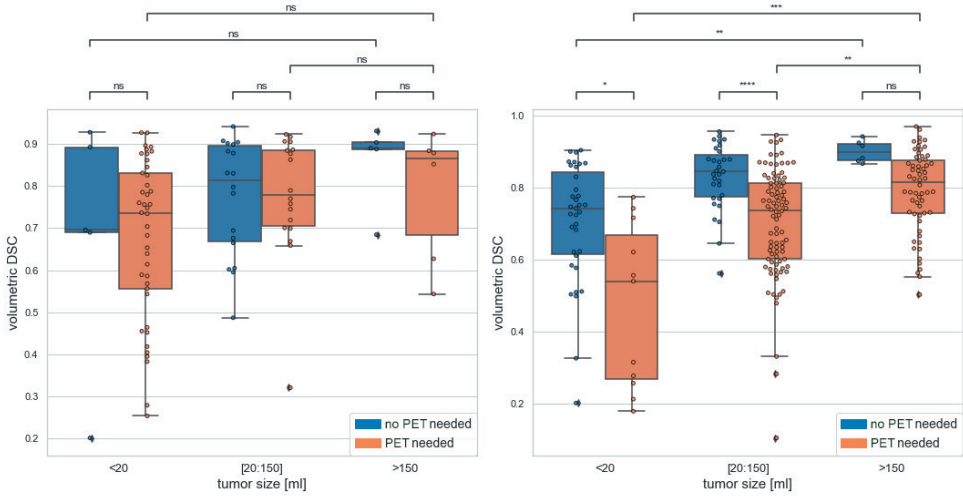


Figure 2. Distributions of DSC performance in sub cohorts grouped based on tumor size and complexity for a) the test dataset and b) the validation dataset.

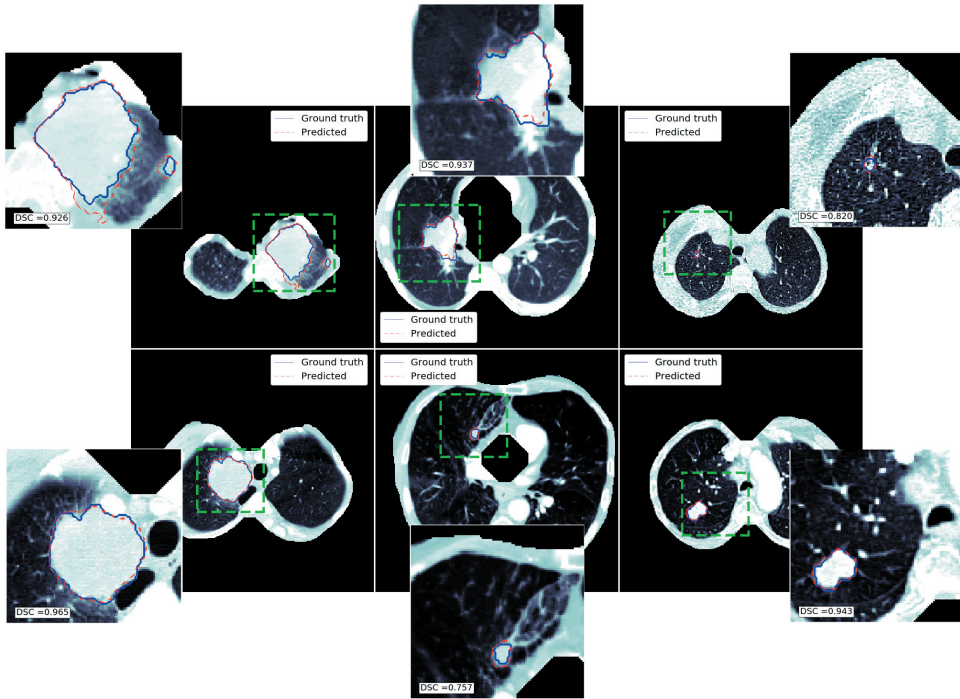


Figure 3. Automatically generated tumour segmentations are shown as red lines while manual segmentations are shown in blue.

Comparison to a published method

A previously published external segmentation model ¹⁹ was evaluated on dataset 8 and compared to our model. The performance of the published model was evaluated using two different inputs: (i) as described in the original article (using patches of 256x256 pixels centered on the tumor); (ii) using the whole slice. For that dataset our method achieved a DSC of 0.88 (IQR = 0.12), whereas the published method achieved a DSC of 0.83 (IQR=0.16) when the cropped tumor regions were used and a DSC of 0.09 (IQR=0.19) in the fully automated configuration (no pre-cropping). Figures for DSC, Ji and H95th are provided in the supplementary materials (Fig. 7 suppl.).

Prognostic power of automatic segmentations

Datasets 1 and 6 were used to compare the prognostic power of measurements extracted from automatically generated and manual contours, as they had available survival data. We calculated the RECIST score and the tumor volume for both the expert and the automatic segmentations, and found that for both metrics the automatically generated segmentations have more prognostic power. Statistical differences in the probability of survival for two groups separated by the median values of these measurements for automated and manual segmentations are reported in Table 4. Kaplan-Meier curves for each method can be found in the supplementary materials (Figures 8, 9 suppl.).

Table 4. Statistical difference between survival groups separated by the median values of RECIST and tumor volume.

Data, (# of patients)	RECIST manual segmentation (p-value)	RECIST automatic segmentation (p-value)	Tumor volume manual segmentation (p-value)	Tumor volume automatic segmentation (p-value)
1, 419	0.00048	< 0.0001	0.00089	< 0.0001
6, 137	0.0038	0.0031	0.031	0.013

In Silico clinical trial

A registered “in silico” clinical trial was performed to assess the following endpoints: 1) the time needed for the processes of manual and automated segmentation; 2) inter and intra-observer variability; 3) preference of experts for manual or automatically generated segmentations.

For the first and second endpoints, seven medical imaging specialists experienced in NSCLC contouring were asked to contour the tumors of 25 patients from dataset 3 while being timed. Our automated method was significantly faster than the fastest participant ($p < 0.0001$). The mean time for the automated method was 2.77 sec/patient (SD = 0.44), whereas the mean time for manual segmentation was 172.19 sec/patient (SD = 158.98) (Fig. 4a).

The mean DSC for intra-observer variability among all experts was 0.86 (IQR=0.13) whereas automated segmentations were 100% reproducible. Individual intra-observer variability scores are reported in figure 4b and the JI and H95th are reported in the supplementary materials (Fig. 10 a, b suppl.). The mean DSC for inter-observer variability was 0.81 (IQR=0.24) (see Fig. 11 suppl.).

The results for assessment of the variability between expert clinicians and the proposed automatic segmentation method achieved on the validation dataset 8 are presented in Figure 5. Our method achieved an average DSC of 0.82 (IQR = 0.14), whereas the average DSC of experts inter-variability was 0.83 (IQR = 0.12).

For the third endpoint, we had 40 participants from 4 different backgrounds: 4 health/medicine master students, 17 computer scientists, 12 medical doctors working in the field of medical imaging, and 7 medical specialists (radiologists or radiation oncologists). In order to quantitatively evaluate the qualitative preferences of experts regarding automated vs manual contours, we developed a software tool which allowed experts to visually compare the segmentations and choose their preference (<https://www.predictcancer.ai/Main.php?page=nsclc-clinical-trial>).

On average, the participants preferred the automatic segmentation above the expert’s contour in 55% (IQR=12%) of the cases (Fig. 6a). Among the groups the qualitative preference scores were as follows: students = 51% (IQR=4%) computer scientists = 52% (IQR=14%), medical doctors = 56% (IQR =12%) and radiologists & radiation oncologists = 59 % (IQR =13%) (Fig. 6b).

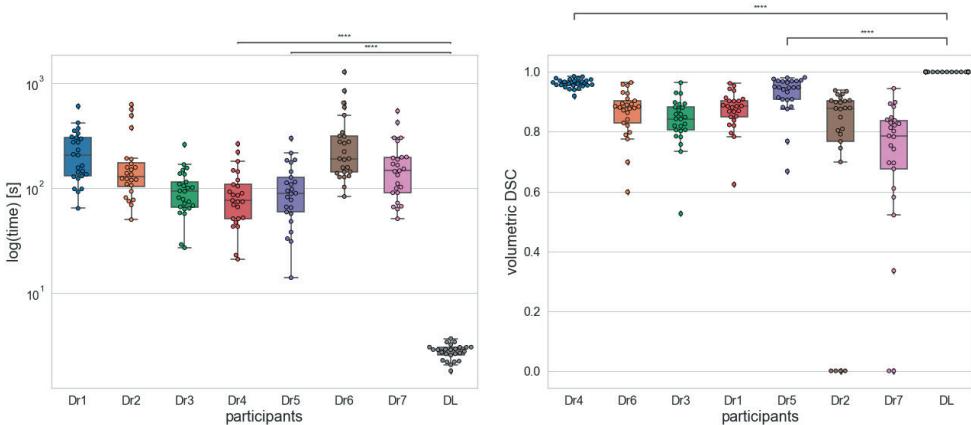


Figure 4. a) Distribution of contouring time for participants and the automated method; b) Volumetric dice similarity coefficient across comparison pairs. Dr1, Dr2, Dr3, Dr4, Dr5, Dr6, Dr7 - represent contours made by the medical doctors, DL -represents automatically generated contours.

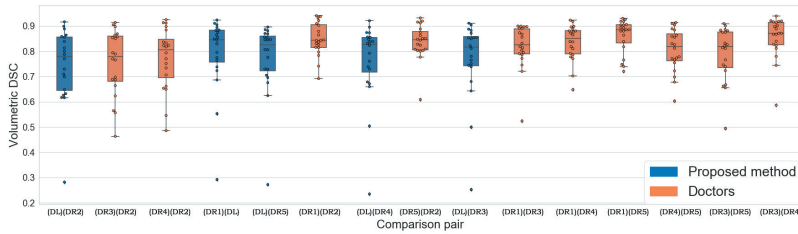


Figure 5. Comparison of volumetric dice similarity coefficient across comparison pairs. DR1, DR2, DR3, DR4, DR5 - represent contours made by the doctors (expert clinicians), DL -represents automatically generated contours.

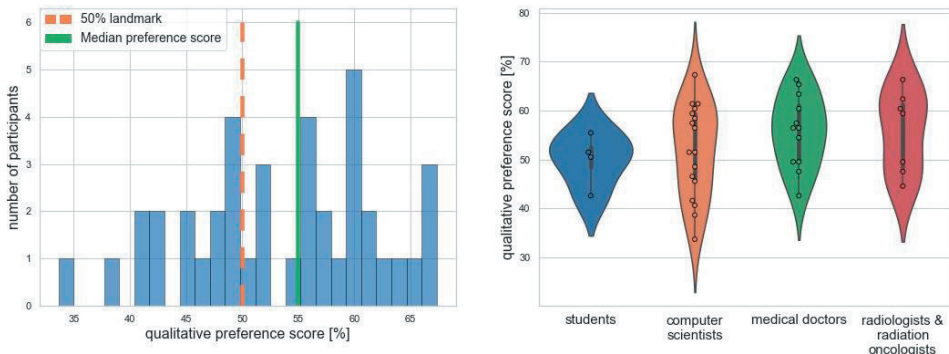


Figure 6. Qualitative preference score: a) distribution of the scores for all participants, b) grouped scores

Discussion

We presented a deep learning-based approach that is able to achieve state-of-the-art detection and 3D volumetric segmentation of NSCLC on CT scans. Although several attempts to develop lung cancer CT detection and segmentation methods have been previously made, our work is novel compared to published solutions, especially in its external validation and ability to work on full thoracic CT scans without further input needed by a human operator. To improve detection and segmentation performance, we introduced several novel steps to the automatic segmentation pipeline: 1) a harmonization routine for the pre-processing of CT scans in order to more comprehensively unify patterns on the images for the models to learn from; 2) a robust computer vision based method to isolate the lung area, allowing the subsequent deep learning step to focus on the region of interest; 3) a dynamically changing loss function for the training procedure, allowing us to control and modify the quality of produced segmentations; 4) CTs of lung abnormalities other than NSCLC were included in the training dataset as negative examples, allowing our method to exclude them from the detection and segmentation process; 5) lung CT slices without contours were also used in the training process as negative samples, thereby increasing the number of unique training samples

and decreasing the false positive rate of the model; 6) although a 2D DL architecture was employed, a 3D post processing routine produced volumetric segmentations. A clinical trial showed that the performance of the automatic segmentation model is acceptable by modern clinical standards and that participants preferred automatic segmentations more often than the manual contours. Furthermore, RECIST and tumor volume based on the automatic contours were able to generate a more significant split of survival groups than manual contours.

To set our model in the context of similar published work, Kamal et al. (2018)¹⁷ used a Recurrent 3D-DenseUNet architecture to segment lung cancers with which they allowed them to obtain a DSC of 0.74 on a validation dataset of 40 patients, compared to our DSC of 0.77 on a validation dataset of 236 patients. Jue et al. (2019)¹⁹ evaluated several 2D convolutional neural network (CNN) architectures such as U-net, Segnet, full resolution residual neural network (FRRN) and incremental multiple resolution residual network (MRRN) to segment patches of 160x160 pixels centered around tumor, achieving DSC of 0.68 on the external validation dataset. Zhang et al. (2020)²⁵ used a modified version of ResNet to automatically segment GTV and achieved an averaged dice similarity coefficient (DSC) of 0.73 on the test set, lacking however external validation of the model. Ardila et al. (2019)¹⁶ developed a deep learning based software which can detect lung cancer on low dose CTs with an AUC of 94.4%. In our study we were not able to evaluate a patient based AUC for lung cancer detection, since all patients had cancer, instead we have demonstrated that our model was able to detect slices containing lung cancer on low dose CTs with a robust AUC of 0.90 both in the test and validation datasets. Additionally, we evaluated the performance of a published 3D U-net based approach on our validation dataset, where our model outperformed the published method.

The state-of-the-art detection accuracy and the fact that it accepts any CT containing the lungs as input means the software can be used as a method for screening and detection of lung cancer. This is further corroborated by the fact that CT scans acquired using different parameters can be directly put in, making our method multi-vendor and multi-reconstruction compliant. The inclusion of cases that were hard to segment without a co-registered PET scan allows the deep learning networks to learn how to differentiate tumors from other lung abnormalities such as atelectasis and tumors with mediastinal involvement, which in conjunction with the accurate segmentation of the 3D tumor volume means it can be used clinically in radiotherapy settings or for big data radiomics (and potentially other) research. The robust automatic volumetric and RECIST measurements will subsequently have a positive impact on sample size calculations for clinical trials²⁶.

Although we attempted to address the flaws and limitations of previous research while developing our software, there were limitations to our work. The ground truth segmentations were originally made on primary NSCLC. Therefore, although the software has a high detection accuracy, it is hypothetically limited to detection and segmentation of primary NSCLC tumors. Moreover, by considering medical expert contours as the ground truth and taking into account the high inter-observer variability of the contouring process²⁷, the deep learning network was also learning inaccuracies, such as contoured air (that certainly is not cancerous). However, this effect can be alleviated by increasing the training dataset size.

In future work we will utilize the evaluated image factors (slice thickness, complexity class, predicted tumor size) in order to give a confidence score to each segmentation produced, providing added information to the user about which segmentations might need more attention.

Further tuning of the model on NSCLC CT scans, and other independent NSCLC datasets can improve the performance of the software, and advance it towards clinical implementation.

The ability of the software developed in this study to handle full thoracic CT scans with different acquisition and reconstruction parameters and without further human intervention represents the pillar for its clinical transition. Clinical application of this software following prospective validation can have a positive impact on the management of lung cancer patients, as it will improve the detection accuracy, and provide a fast, consistent and reliable volumetric segmentation for treatment (evaluation) purposes. Furthermore, the use of the software in large radiomics studies will allow automation and will reduce the time needed to complete the studies in a robust manner, as it will significantly decrease the time needed for the rate-limiting part of the workflow - tumor segmentation.

Methods

Description of data

The CT scans of 1343 NSCLC patients were retrospectively collected and anonymized by each center and approved by the respective institutional review boards. In this study, which followed the Standards for Reporting of Diagnostic Accuracy Studies statement²⁸, the requirement for written informed consent was waived. The images in dataset 8 were segmented by five radiation oncologists, which allowed us to compare the performance of the deep learning segmentation model to multiple manual delineations. All other segmentations were performed by a radiologist or radiation oncologist at the center where diagnosis was made, and checked by at least one segmentation expert at our site.

The expert segmentations were considered as the ground truth for training and further evaluations. Fifteen patients from various datasets were excluded due to missing tumor contours and the lack of a PET scan to perform the segmentations according to clinical protocol. Survival data and CT scans for datasets 1 and 6 were collected from the open sources.

Image pre-processing

Data inhomogeneity necessitated the harmonization of CT data in order to achieve comparable representations of the tumor region. Furthermore, several steps were introduced to reduce computational power requirements and image noise, and to optimize the contrast. The first step is the extraction of a 3D array with voxel intensity values represented as Hounsfield Units (HU) from Digital Imaging and Communications in Medicine (DICOM) data. Next, the image contrast is enhanced using a lung window setting (window width (WW) of 1500HU and window level (WL) of -600 HU) to highlight lung structures. All voxel intensities outside of the upper and lower limits are assigned the value of the closest limit. Following this, nearest neighbor interpolation is applied to obtain isotropic spatial resolution in the axial plane so that each pixel has a size of $1 \times 1 \text{ mm}^2$. After spatial normalization, an image with standard bone window settings (WW: 1800, WL: 400) is saved, as it is used as an input in the lung isolation step of the workflow. In order to smooth the effect of different reconstruction methods on the image and to reduce computational burden, intensity values are aggregated into bins of equal width. This also allows optimization of storage and image processing by packing the images into a much shorter 8-bit integer range and by filtering high frequency noise. Hereafter, the image is cropped or padded with air intensity values to arrive at a resolution of 512×512 pixels, which is chosen as input for the selected deep learning architecture. All image processing and deep learning modelling steps were performed in Python 3.7 with the libraries and respective versions detailed in supplementary materials Table suppl. 1.

Lung region isolation

A robust algorithm for the isolation of the lung region was developed in order to focus on the ROI and allow for the use of whole body CT scans as input. First, the CT couch is detected and removed from the image volume. Air-filled connected volumes are detected and region growing and morphological operations are applied in order to remove small vessels and to connect adjacent regions, resulting in a 3D binary lung mask. The spine axis is identified and the lung mask is halved and symmetrically flipped about the sagittal plane, keeping the union of the flipped and the original lung masks. By doing so, the algorithm is optimized for handling lung abnormalities such as atelectasis, pulmonary infiltration, consolidation, and fibrosis. To accurately identify the spine axis, a further algorithm was developed which identifies the center of the spine

using the stored preprocessed image with bone window settings as described in the previous section (Fig. 12 a suppl.). A “bone image” slice containing the lung is projected onto the coronal plane and filtered with a seventh order moving average filter (Fig. 12 b-c suppl.). This is repeated for the first five slices in which the lung mask is present in order to find a starting point for the center spine position S_0 . The axis of the spine is positioned normally to this point (Fig. 12 d suppl.).

$$S_0 = \frac{1}{n} \sum_{z=0}^n P_z \quad (\text{Equation 1})$$

Where P is a central spine point for the current axial slice, n is the number of slices ($= 5$). Due to irregularities of patient positioning and anatomy, the central spine position S_t is recalculated slice-wise by using exponential smoothing:

$$S_t = \alpha \cdot x_t + (1-\alpha) \cdot S_{t-1} \quad (\text{Equation 2})$$

Where x is a central spine point based on the filtered signal for the current axial slice, and α is the weighting coefficient ($= 0.3$).

This method of flipping the lung mask allows for the inclusion of regions that contain large-sized abnormalities, such as lung collapse, which obscure parts of the lung, whereas commonly used methods exclude those regions (Fig. 12 f-g suppl.).

A morphological dilation with the circle kernel ($r=5$) is applied to the resulting lung mask in order to have a margin around the lung area. The final binary lung mask is used to isolate the lung region within the original image by setting all the voxel values outside the mask to the normalized air value.

Tumor detection and segmentation

The widely used 2D U-net convolutional neural network (CNN) was employed for slice-wise tumor segmentation³⁰⁻³³. The axial projection was used to train the network due to the higher resolution of image representation in this plane. To improve segmentation performance, several changes were made to the original CNN architecture. First, rectified linear unit (ReLU) activations were replaced with Exponential Linear Unit (ELU) in order to alleviate the gradient vanishing problem and kick-start the training process³⁴. Second, to capture deeper features from the CT scan, an additional convolutional block was added to the U-net encoder so that the smallest analyzed path resolution is 16x16 pixels. In addition, dropout layers with the dropout rate ($p = 0.5$) were introduced prior to the 3 last layers of U-net encoder to prevent overfitting³⁵.

A 2D CNN architecture was chosen for several reasons: 1) by using a 2D input the

training dataset can be increased by more than a factor of 60, as overall more than 60000 unique slices were available in the training set; 2) due to calculation costs, most present deep 3D architectures could analyze only a sub volume of the medical image^{36,37}, or they require a dimensionality reduction using interpolation or other image processing methods. 2D architectures do not have this problem and can process CT scans in the original resolution; 3) our main goal was to develop a pipeline that can be used in a clinical setting, and a 2D architecture allows for significantly lower requirements for executing PC. Our software does not require GPUs and can run on a regular laptop (Intel Core i5, 2.5GHz, 8GB RAM).

In order to increase robustness of the system to a wide range of imaging parameters, the training dataset was expanded using augmentation techniques with the following parameters: random rotation around the image center pixel in a range of 0-25 degrees with a probability of 60%, random horizontal and vertical shifts of the image in the range of 12% of image shape with a probability of 25%, random zooming of the image with a maximum of 3% of the image shape with a probability of 10%.

The loss function was calculated by combining the Dice similarity coefficient (DSC) loss and the binary cross-entropy, and privilege was given to the DSC loss during the first 50 epochs. The privilege was defined by the coefficients before the DSC and cross-entropy terms in the loss function. By adding the binary cross-entropy component to the loss function, negative samples (slices without contour) could also contribute to the training.

The model was trained for 300 epochs using eight NVIDIA GTX 1080 Ti GPUs. The Adam algorithm was used for the stochastic optimization of the loss function³⁸. The cosine annealing scheduler was used to adjust the learning rate during the training process. A checkpoint function tracking the DSC on the test dataset was used to keep the best weights.

Predicted 2D binary masks are stacked into a 3D volume and connected component extraction is applied as a post-processing step, whereby only spatially connected mask regions are extracted³⁹. The final mask is resampled to the original image shape using nearest neighbor interpolation.

Evaluation metrics

The ability of the system to detect tumors was assessed by calculating the slice-wise accuracy and generating a confusion matrix. Slices without segmentation were considered as not containing tumor tissue. Automatically generated binary masks were resampled to the original image resolution using nearest neighbor interpolation before comparing

with manual segmentations. The contouring performance of the proposed pipeline, as well as the doctors variability, were assessed by using the volumetric Dice similarity coefficient (DSC), Jaccard index (Ji) and 95th percentile Hausdorff distance (H95th). The DSC is a measure of overlap between two volumes and was computed as:

$$DSC = \frac{2 \cdot |F \cap G|}{|F| + |G|} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (\text{Equation 3})$$

Jaccard index, used for gauging the similarity between two volumes, was computed as:

$$Ji = \frac{|F \cap G|}{|F \cup G|} = \frac{TP}{TP + FP + FN} \quad (\text{Equation 4})$$

where F and G are the sets of voxels corresponding to the ground truth and the automatic segmentation, respectively. TP is the number of true positive voxels, FP is the number of false positive voxels and FN is the number of false negative voxels.

To evaluate the maximum deviation between the automatic segmented surface boundary and the ground truth surface boundary, the 95th percentile of Hausdorff distance (H95th) was used. Hausdorff distance (H) is defined as:

$$H(A, B) = \max\left\{ \sup_{a \in S_a} \inf_{b \in S_b} d(a, b), \sup_{b \in S_b} \inf_{a \in S_a} d(b, a) \right\} \quad (\text{Equation 5})$$

where a and b are the points on the voxel sets A and B, which represent the ground truth and the automatic segmentation, respectively. S_a and S_b are the surfaces of A and B.

In addition to the model performance evaluation on the test and validation datasets, the variability between expert clinicians was assessed and displayed against the performance of our method by comparing the volumetric DSC among all possible comparison pairs, i.e. experts were compared with each other as well as with the proposed method.

To better gauge the performance of our model under varying circumstances it was evaluated with regard to slice-thickness, tumor complexity, and tumor size. Tumor size sub groups were chosen based on the overall tumor size distribution in the training set. Furthermore, expert subjective tumor complexity labels were defined. To describe the complexity of the tumor, two medical doctors were asked to label the test and validation dataset as follows: for tumors where segmentation cannot be performed without a corresponding PET scan the labels were set to “1”, and “0” otherwise. In case of disagreement, the label “1” was chosen.

Statistical analysis

For all non-normally distributed scores the median and interquartile range (IQR) were reported, as well as the frequency histograms²⁹. Statistical significance was assessed using a two-sided Mann-Whitney-Wilcoxon test with Bonferroni correction. Survival evaluation was done in R (version 4.0.2) using survival (version 3.1-12) and survminer (version 0.4.7) packages. To estimate the difference between survival groups a log-rank test was applied. High and low survival groups were separated by the median tumor volume or median RECIST measurement respectively. A random sampling with replacement bootstrapping strategy was used to compute confidence intervals for AUC values.

An *In Silico* clinical trial

This trial was registered at clinicaltrials.gov (NIH: NCT04164186). For the first and second endpoints (the time needed for the processes of manual and automated segmentation, and inter and intra-observer variability), participants used a state of the art commercial software (MIM version 7.0.4) to produce the segmentations. In order to make the conditions of the trial close to the real clinical practice, experts had CT and PET scans available for each patient and they were able to use a semi-automated segmentation solution provided by MIM, while the proposed method generated the segmentation using only CT scans.

For the third endpoint (preference of experts for manual or automatically generated segmentations), a software tool was developed in-house. The tool has two interactive screens with the first screen showing the description of the experiment and a small questionnaire. In order to analyse preferences at different levels of expertise, the participants were asked to specify their training (e.g. radiologist, radiation-oncologist, medical doctor). The second screen displays comparisons between pairs of segmented axial CT slices (automatic vs. expert) with randomized screen positions, blinded to the participant. For each comparison pair, the participants were asked to select the more accurate contour. Finally, a table was generated containing the choices made. Screenshots of this tool are provided in supplementary materials (Fig. 13-14 suppl.).

The software tool presents scans and contours from the external validation datasets 8. It randomly selects 100 pairs of contoured CT slices, where the DSC between the contours was higher than 0.7. During the assessment, participants were able to adjust the image contrast by changing window settings (WW and WL), and to leave comments.

The preference of the experts was evaluated using the qualitative preference score, defined as:

$$Score_{method} = \frac{n_{method}}{n_{overall}} \times 100\% \quad (\text{Equation 6})$$

where n_{method} is a number of times where preference was given to the proposed method, $n_{overall}$ is a number of cases in total.

Data availability

Philippe Lambin and Henry Woodruff should be addressed for correspondence and material requests.

The open source data used in this article has the appropriate references. The private data sets used in this article are available from the corresponding author upon request subject to ethical review.

Code availability

Code for the article will be available on the github upon publication.

Clinical trial app is available by the following link: <https://www.predictcancer.ai/Main.php?page=nsclc-clinical-trial>.

References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* vol. 68 394–424 (2018).
2. Postmus, P. E. *et al.* Early and locally advanced non-small-cell lung cancer (NSCLC): ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* **28**, iv1–iv21 (2017).
3. Jaffray, D. A. Image-guided radiotherapy: from current concept to future perspectives. *Nat. Rev. Clin. Oncol.* **9**, 688–699 (2012).
4. Barrett, A., Dobbs, J. & Roques, T. *Practical Radiotherapy Planning Fourth Edition*. (CRC Press, 2009).
5. Stroom, J. C. & Heijmen, B. J. M. Geometrical uncertainties, radiotherapy planning margins, and the ICRU-62 report. *Radiother. Oncol.* **64**, 75–83 (2002).
6. Wolchok, J. D. *et al.* Guidelines for the Evaluation of Immune Therapy Activity in Solid Tumors: Immune-Related Response Criteria. *Clinical Cancer Research* vol. 15 7412–7420 (2009).
7. Erasmus, J. J. *et al.* Interobserver and Intraobserver Variability in Measurement of Non-Small-Cell Carcinoma Lung Lesions: Implications for Assessment of Tumor Response. *J. Clin. Oncol.* **21**, 2574–2582 (2003).
8. Lambin, P. *et al.* Radiomics: extracting more information from medical images using advanced feature analysis. *Eur. J. Cancer* **48**, 441–446 (2012).
9. Lambin, P. *et al.* Radiomics: the bridge between medical imaging and personalized medicine. *Nat. Rev. Clin. Oncol.* **14**, 749–762 (2017).
10. Aerts, H. J. W. L. *et al.* Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.* **5**, 4006 (2014).
11. Kumar, V. *et al.* Radiomics: the process and the challenges. *Magn. Reson. Imaging* **30**, 1234–1248 (2012).
12. Ibrahim, A. *et al.* Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. *Seminars in Nuclear Medicine* (2019) doi:10.1053/j.semnuclmed.2019.06.005.
13. Kalmet, P. H. S. *et al.* Deep learning in fracture detection: a narrative review. *Acta Orthop.* **91**, 362 (2020).
14. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science* 234–241 (2015) doi:10.1007/978-3-319-24574-4_28.
15. Szegedy, C. *et al.* Going Deeper with Convolutions. *arXiv [cs.CV]* (2014).
16. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
17. Kamal, U., Rafi, A. M., Hoque, R. & Hasan, M. K. Lung Cancer Tumor Region Segmentation Using Recurrent 3D-DenseUNet. *arXiv [eess.IV]* (2018).
18. Ray, A. *Lung Tumor Segmentation via Fully Convolutional Neural Networks*. (2016).

Validated fully automated detection and segmentation of non-small cell lung cancer

19. Jiang, J. *et al.* Multiple Resolution Residually Connected Feature Streams for Automatic Lung Tumor Segmentation From CT Images. *IEEE Trans. Med. Imaging* **38**, 134–144 (2019).
20. Mackin, D. *et al.* Measuring Computed Tomography Scanner Variability of Radiomics Features. *Invest. Radiol.* **50**, 757–765 (2015).
21. Aerts, H. *et al.* Data from NSCLC-radiomics. *Cancer Imaging Archive* (2015).
22. Bakr, S. *et al.* A radiogenomic dataset of non-small cell lung cancer. *Scientific Data* vol. 5 180202 (2018).
23. Clark, K. *et al.* The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* **26**, 1045–1057 (2013).
24. van Baardwijk, A. *et al.* PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. *Int. J. Radiat. Oncol. Biol. Phys.* **68**, 771–778 (2007).
25. Zhang, F., Wang, Q. & Li, H. Automatic Segmentation of the Gross Target Volume in Non-Small Cell Lung Cancer Using a Modified Version of ResNet. *Technol. Cancer Res. Treat.* **19**, 1533033820947484 (2020).
26. Revel, M.-P. *et al.* Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? *Radiology* **231**, 453–458 (2004).
27. Velazquez, E. R. *et al.* A semiautomatic CT-based ensemble segmentation of lung tumors: Comparison with oncologists' delineations and with the surgical specimen. *Radiotherapy and Oncology* vol. 105 167–173 (2012).
28. Cohen, J. F. *et al.* STARD 2015 guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open* **6**, e012799 (2016).
29. Habibzadeh, F. How to report the results of public health research. *J Public Health Emerg* **1**, 90–90 (2017).
30. Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T. & Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016* 424–432 (2016) doi:10.1007/978-3-319-46723-8_49.
31. Norman, B., Padoia, V. & Majumdar, S. Use of 2D U-Net Convolutional Neural Networks for Automated Cartilage and Meniscus Segmentation of Knee MR Imaging Data to Determine Relaxometry and Morphometry. *Radiology* **288**, 177–185 (2018).
32. Livne, M. *et al.* A U-Net Deep Learning Framework for High Performance Vessel Segmentation in Patients With Cerebrovascular Disease. *Front. Neurosci.* **13**, 97 (2019).
33. Hashimoto, F., Kakimoto, A., Ota, N., Ito, S. & Nishizawa, S. Automated segmentation of 2D low-dose CT images of the psoas-major muscle using deep convolutional neural networks. *Radiol. Phys. Technol.* (2019) doi:10.1007/s12194-019-00512-y.
34. Clevert, D.-A., Unterthiner, T. & Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *arXiv [cs.LG]* (2015).
35. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
36. Yu, L., Yang, X., Chen, H., Qin, J. & Heng, P.-A. Volumetric ConvNets with mixed residual

- connections for automated prostate segmentation from 3D MR images. in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* 66–72 (AAAI Press, 2017).
37. Wu, W. *et al.* Segmentation of pulmonary nodules in CT images based on 3D-UNET combined with three-dimensional conditional random field optimization. *Medical Physics* vol. 47 4054–4063 (2020).
 38. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *arXiv [cs.LG]* (2014).
 39. Dillencourt, M. B., Samet, H. & Tamminen, M. A general approach to connected-component labeling for arbitrary image representations. *Journal of the ACM* vol. 39 253–280 (1992).
 40. Zhou, S., Cheng, Y. & Tamura, S. Automated lung segmentation and smoothing techniques for inclusion of juxtapleural nodules and pulmonary vessels on chest CT images. *Biomed. Signal Process. Control* **13**, 62–70 (2014).
 41. Shojaii, R., Alirezaie, J. & Babyn, P. Automatic lung segmentation in CT images using watershed transform. in *IEEE International Conference on Image Processing 2005* vol. 2 II–1270 (2005).

Acknowledgements

Sergey Primakov, Manon Beauque and Iva Halilaj acknowledge the financial support of Marie Skłodowska-Curie grant (PREDICT - ITN - No. 766276).

Authors acknowledge financial support from ERC advanced grant (ERC-ADG-2015 n° 694812 - Hypoximmuno), ERC-2018-PoC: 813200-CL-IO, ERC-2020-PoC: 957565-AUTO.DISTINCT, Authors also acknowledge financial support from SME Phase 2 (RAIL n°673780), EUROSTARS (DART, DECIDE, COMPACT-12053), the European Union's Horizon 2020 research and innovation programme under grant agreement: ImmunoSABR n° 733008, FETOPEN- SCANnTREAT n° 899549, CHAIMELEON n° 952172, EuCanImage n° 952103, TRANSCAN Joint Transnational Call 2016 (JTC2016 CLEARLY n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine (EURADIOMICS n° EMR4). This work was supported by the Dutch Cancer Society (KWF Kankerbestrijding), Project number 12085/2018–2.

Author contributions

S.P, P. L., A. J., A. I., H. C. W. conceived the idea of the article. M. B, S. K., E. K., A. I., S. S., I. H., J. W., R. M., H. G., L. H., O. M., M. S., R. G., G. W. participated in the data acquisition and clinical trial. S.P. implemented the analysis. J. V. T., A. J., A. I., H. C. W., P. L., S. K., R. G. contributed to the writing of the manuscript. H.C.W, A.J., P.L supervised the work.

H.C.W. approved the submitted version and has agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

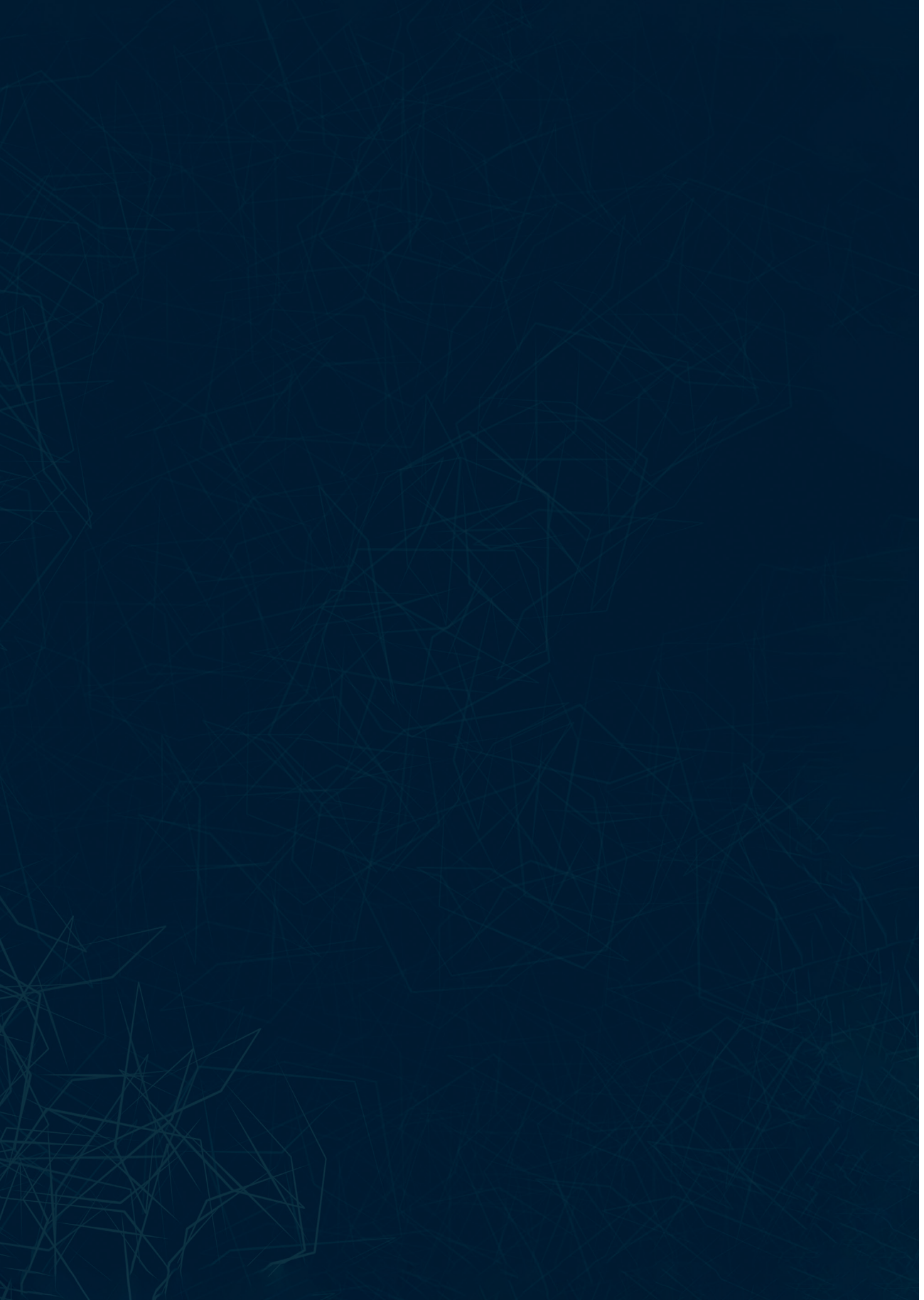
Competing interests

Dr. Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Oncoradiomics, ptTheragnostic, Health Innovation Ventures and DualTpharma. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in-kind manpower contribution from Oncoradiomics, BHV, Merck and Convert pharmaceuticals. Dr Lambin has shares in the company Oncoradiomics SA and Convert pharmaceuticals SA and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patentable invention (software) licensed to ptTheragnostic/ DNAmito,

Oncoradiomics and Health Innovation Ventures.

Dr. Woodruff reports, outside of current manuscript, (minority) shares in the company Oncoradiomics.

Dr. Lizza Hendriks reports, none related to current manuscript, outside of current manuscript: research funding Roche Genentech, Boehringer Ingelheim, AstraZeneca (all institution); advisory board: Boehringer, BMS, Eli Lilly, Roche Genentech, Pfizer, Takeda, MSD, Boehringer Ingelheim, Amgen (all institution); speaker: MSD (institution); travel/conference reimbursement: Roche Genentech (self); mentorship program with key opinion leaders: funded by AstraZeneca; fees for educational webinars: Quadia (self); interview sessions funded by Roche Genentech (institution); local PI of clinical trials: AstraZeneca, Novartis, BMS, MSD /Merck, GSK, Takeda, Blueprint Medicines, Roche Genentech, Janssen Pharmaceuticals, Mirati



PART V



14

Chapter 14

General discussion
and future perspectives

Medical image analysis (MIA) using artificial intelligence (AI) methods has become a widely researched topic. AI tools can become efficient, reliable and none-(minimally) invasive clinical decision support systems with the potential to improve personalized care [1,2]. To date, many studies investigated and reported on the potential of different AI applications for MIA. A significant number of studies applied handcrafted radiomics or deep learning methods to perform different tasks on different medical imaging modalities [3,4].

In this thesis, the focus was on two of the AI methods: (i) handcrafted radiomics; and (ii) deep learning (DL). For handcrafted radiomics, the overarching aim was to gain deeper insights into the reproducibility of handcrafted radiomics features (HRFs), in order to develop objective methods addressing the issue. For DL, the overarching aim was to investigate novel applications in MIA. In this chapter, an extended discussion of the work done in this thesis, as well as future perspectives, are provided.

Evaluation of the conventional handcrafted radiomics workflow

The conventional handcrafted radiomics workflow includes imaging data collection, lesions segmentation, HRFs extraction, signature development. Each of these steps are faced by some challenges that can affect the reproducibility of extracted HRFs significantly [5,6]. In Chapter 2, a literature search was performed. Based on the findings, and previous experiments, a radiomics workflow that takes into consideration the reproducibility of HRFs was proposed. The workflow consists of several steps, with the addition of a reproducibility analysis step. The added step introduces the need for the assessment of the repeatability, reproducibility and harmonizability of HRFs based on the data being analyzed. The aim is to guide the development of generalizable radiomic signatures that could be used across scans acquired with different imaging parameters.

In Chapter 3, the conventional radiomics workflow was applied on two independent MRI datasets to predict complete pathologic response to treatment in breast cancer patients. The scans were acquired using different imaging vendors and imaging parameters across the datasets. The robustness of the signatures developed was assessed by performing the radiomics analysis 100 times, with random splits of the training and testing datasets. The analysis showed that different HRFs were selected with a wide range of performance values, indicating the lack of robustness of the signatures.

In Chapter 4, a similar analysis to that in Chapter 3 was performed on axillary MRI scans to predict node status in breast cancer patients. The study included a smaller number of scans that were acquired using different imaging parameters. In concordance with the findings reported in Chapter 3, the results indicated the signatures lacked robustness. In addition to the literature, results of Chapters 3 and 4 indicated the need

for investigating the reproducibility of HRFs in different scenarios, and the need for methods to assess the reproducibility of and harmonize HRFs that are extracted from scans acquired with different imaging parameters. Chapters 3 and 4 confirmed the hypothesis that the assessment of the reproducibility of HRFs must be a corner stone in radiomics analyses.

Investigations of the reproducibility of HRFs, and validation of the proposed framework

Chapter 5 presented a thorough literature search into the currently used harmonization methods in radiomics analyses. Based on the findings, there is currently a need for harmonization methods that are specifically designed to incorporate the effects of various imaging parameters on HRFs, while also considering the different levels of complexities of different HRF groups.

In Chapter 6, the robust radiomics analysis workflow proposed in Chapter 2 was applied on a set of 13 phantom CT scans. The scans were acquired using different imaging vendors and parameters. HRFs were extracted from these scans and their reproducibility was assessed using the concordance correlation coefficient (CCC). The result showed that only a small number of HRFs were insensitive to the variations in the imaging parameters analyzed. The reproducibility of the majority of HRFs was dependent on the variations in imaging parameters. A given HRF could be reproducible in some scenarios, but not the others. Furthermore, the performance of ComBat harmonization was assessed on the phantom scans by calculating the CCC. The results indicated that the ability of ComBat to harmonize a given HRF is dependent on the variations in imaging parameters. These findings supported the use of the proposed radiomics workflow in radiomics studies analyzing scans that were acquired with different imaging parameters. The findings are also in concordance with previous studies that investigated the reproducibility of HRFs in different scenarios [6,7].

Chapter 7 included an experiment to assess the effects of variations in pixel spacing while all other imaging parameters were fixed. The workflow proposed in chapter 2 was modified to accommodate harmonization techniques other than ComBat. The data analyzed included two sets of CT scans of a 10-layer phantom. Each set consisted of 7 CT scans that were acquired with the same imaging parameters except for the pixel spacing. Each layer was segmented as a volume of interest (VOI), resulting in 10 VOIs per scan. The reproducibility of HRFs across pairs of phantom CTs was assessed using the CCC. Findings indicated that some HRFs are insensitive to differences in pixel spacing, while the reproducibility of the remaining HRFs was dependent on the degree of variation in pixel spacing. In addition, the effects of ten different image resampling methods, and

ComBat harmonization on HRFs extracted from the two sets were assessed. On average, scans resampled using cosine windowed sinc interpolation showed the highest numbers of concordant HRFs compared to other resampling methods. Nevertheless, the effects of image resampling and ComBat harmonization on the reproducibility of HRFs were found to be dependent on the variations in the scans being analyzed. In other words, a given HRF could be harmonizable with image resampling\ComBat in some scenarios but not the others. These results further consolidated the need for reproducibility studies in radiomics analyses including scans acquired with different imaging parameters.

A suggestion that ComBat harmonization must be applied to each layer of the phantom separately was made [8]. Accordingly, each layer of the phantom was subdivided into 16 VOIs, and ComBat harmonization was applied on each of the layers separately. The results of this experiment augmented the findings in Chapter 7, specifically, the need for reproducibility analysis to evaluate both the reproducibility and harmonizability of HRFs across CT scans acquired with different imaging parameters. The slightly modified robust radiomics analysis workflow (figure 1) can be utilized to develop generalizable radiomic signatures. The modification was done during further experiments, to allow the generalization of the workflow to different radiomics harmonization methods.

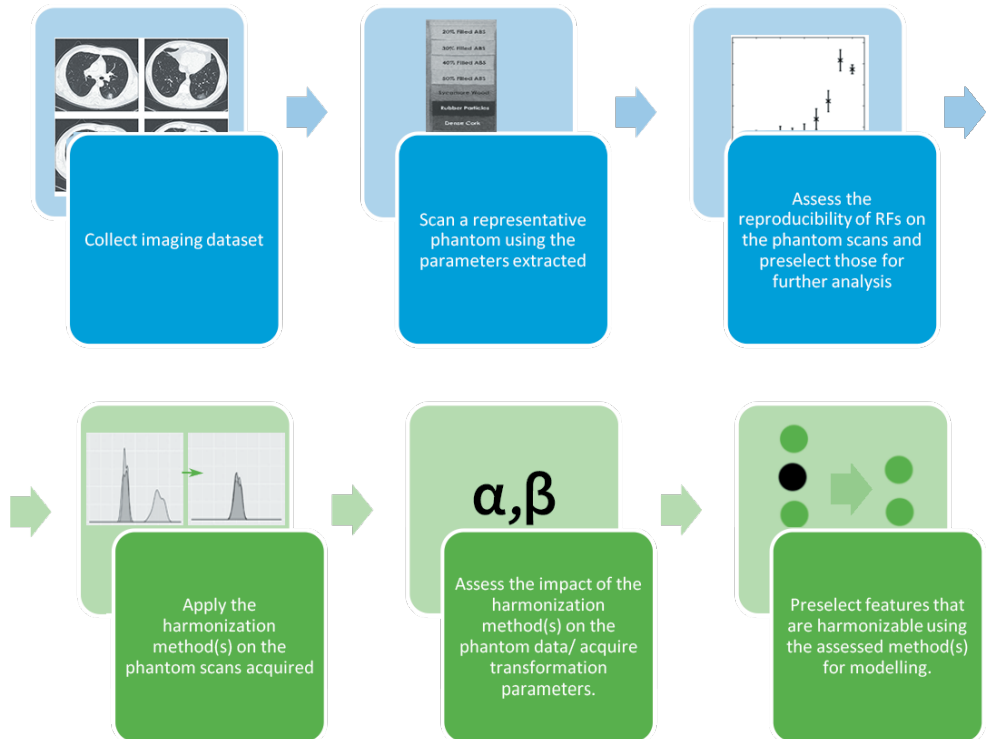


Figure I: The final proposed robust radiomics analysis workflow [12].

In Chapter 8, the reproducibility of hepatocellular carcinoma (HCC) HRFs extracted from different phase contrast enhanced CT scans was assessed. Arterial and portal venous CT scans of HCC patients enrolled in a clinical trial were available for the analysis. The regions of interest (ROIs) were segmented on one of the phases, and copied to the other phase. A fraction of HRFs were found to be reproducible across both phases, when no imaging parameters varied. The potential of ComBat harmonization to remove differences in HRFs values attributed to differences in imaging phase. ComBat harmonization increased the number of reproducible HRFs across phases by 1%. The conclusions derived from this study were that a number of HRFs can be interchangeably used across arterial and venous phase CT scans, and that the combination of these scans per patient could maximize the information extracted from HCC lesions. Furthermore, Chapters 6, 7, 8 and 9 confirmed the hypothesis that HRFs are differently affected by the differences in variations in imaging parameters, and that the performance of different harmonization methods is dependent on the variations in the data being harmonized.

Chapter 9 describes an experiment that was designed to investigate the effects of inter-reader variability on the reproducibility of HRFs extracted from breast MR scans. Four readers with varying experience in medical image segmentation were asked to segment a number of breast MR scans. HRFs were extracted from each of the segmentations, and the reproducibility of HRFs across the multiple segmentations was assessed using the interclass correlation coefficient (ICC). The majority of HRFs (-67%) were found to be significantly affected ($ICC < 0.9$). The findings suggest the need to address inter-reader variability, when the datasets are segmented by different observers. This chapter has also confirmed the hypothesis that HRFs are subject to inter-reader variability, and that attention must be paid to that in future radiomics studies.

Chapter 10 is an MRI test retest experiment to assess the repeatability of breast tissue HRFs. A number of healthy volunteers were scanned on two dates, with nine scans per session. Different MR sequences (T1W, T2W and ADC maps) were analyzed. Furthermore, different image preprocessing techniques were assessed on the different sequences. The repeatability of HRFs varied across the sequences and image preprocessing techniques. The majority of breast MR HRFs were not found to be repeatable. The findings add to the evidence necessitating reproducibility analyses in radiomics studies. This experiment confirmed the hypothesis that some HRFs are not reproducible in test-retest scenarios.

In Chapter 11, the findings from previous chapters led to designing an experiment to assess the reproducibility of HRFs across a wide range of variations in imaging parameters. The analyzed CT phantom scans resulted in 31375 different scenarios. The

reproducibility of HRFs across pairs of scans was assessed using the CCC. The number of reproducible HRFs varied depending on the variations in the imaging parameters. Scans that were acquired with similar convolution kernels, slice thickness and pixel spacing values showed a higher number of reproducible HRFs, compared to scans acquired with big differences in those parameters. ComBat harmonization and image resampling were investigated as potential methods to harmonize HRF values across all the 31375 scenarios. ComBat harmonization resulted on average in a higher number of reproducible HRFs compared to image resampling. Nevertheless, only 1% of HRFs were harmonizable with either of the methods regardless of the variations in imaging parameters. Furthermore, a quantitative score was developed based on the variations in imaging parameters, confirming the hypothesis that a quantitative tool can be developed to assess the reproducibility of HRFs across scans acquired differently. The score can be used to assess the percentage of reproducible HRFs across CT scans acquired with different imaging parameters. Robustness analysis indicated the robustness of the develop score in assessing the reproducibility of HRFs.

Some application of DL on medical images

Chapter 12 describes the development of a DL algorithm that could be used to classify bone scintigraphies whether they contain metastatic bone lesions. Data was collected from three independent medical centers. The model was trained and tested on bone scintigraphies collected at two different centers, and was externally validated on the data from the third center. In addition, the explainability of the developed algorithm was enhanced using gradient-CAM method, which produces activation maps to indicate the regions that resulted in a positive decision. Furthermore, the performance of the model was compared to that of a group of nuclear physicians in an *in silico* trial. The results highlight the potential of DL algorithms to be used as clinical decision support tools.

Chapter 13 describes the development of a CE marked software for the automated detection and segmentation of non-small cell lung cancer on CT scans. The software is DL algorithm that was trained and externally validated on a large number of multicenter datasets. The software includes several consecutive steps that include the harmonization of the CT scans, isolation of the lung region, and segmentation of lung lesion(s). The performance of the software was further assessed in an *in silico* trial, which showed that on average, radiologists and radiation oncologists preferred the automatically generated segmentations. Chapters 12 and 13 confirmed the hypothesis that DL algorithms could potentially perform some clinical tasks with high accuracy in significantly short times.

Future perspectives

During the last decades, there has been an exponential growth in the number of studies investigating the potential use of artificial intelligence techniques in medical image

analysis [10,11]. As with the majority of scientific fields (Figure 2), the development of handcrafted radiomics field has gone through several phases. There was a hype in the expectations for the field, evidenced by the large number of radiomic studies in the literature. Then, there has been a shift of focus from making predictions, to addressing the issues hindering utilization of the full potential of the field, as the field is currently in the “slope of enlightenment” phase. The issues include, but are not limited to, the reproducibility of the quantitative imaging features, the explainability of signatures, and the need for big data.

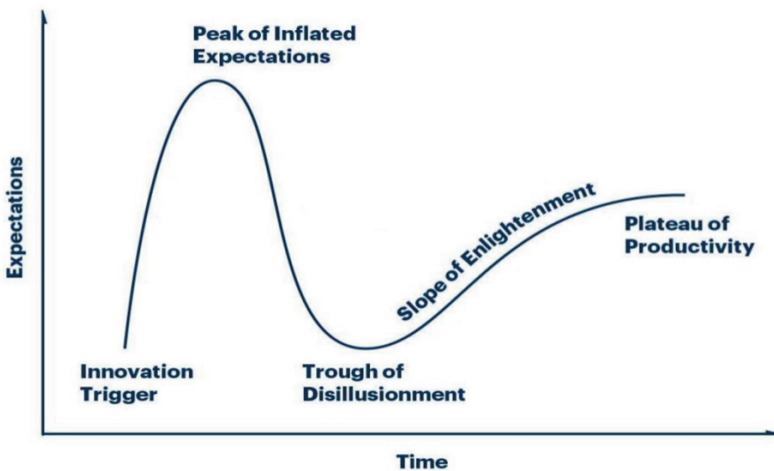


Figure 2. Phases of field development (Credit: Carole Goble).

In this thesis, the majority of the work was focused on understanding and mitigating the limitations of HRFs, potentially participating in the upward progression of the “slope of enlightenment” phase of the field, and hasten the transition to the “plateau of productivity”. Our proposed robust radiomics analysis workflow (Figure 1) would aid the development of generalizable radiomics signatures in future radiomics studies. In addition, the novel MaasPenn radiomics reproducibility score can be utilized as a tool for deciding on scans inclusion, as well as assessing the generalizability of developed radiomics signatures.

Future research into the reproducibility of HRFs across different imaging parameters should include a wider range of imaging parameters than those investigated in Chapter 11. A larger dataset with more variations could enhance our understanding of the combined effects of differences in imaging parameters on the reproducibility of HRFs, and thereby the ability to enhance the performance of the score. In addition, feature specific reproducibility scores could be investigated in future research to further improve the ability to develop robust radiomic signatures, ultimately leading to improved

personalized management and patient outcomes. Furthermore, since the majority of HRFs can be significantly affected by the variations in imaging parameters, there is a current need for a radiomics-specific harmonization method(s) that take(s) into account the effects of differences in imaging parameters, as well as the nature of different HRFs. Last but not least, attention must also be paid to the repeatability of HRFs in test-retest, as well as their sensitivity to inter-reader variability.

Further research into the applications of DL methods on medical images is recommended to assess the effects of differences in imaging parameters on the performance of DL algorithms, to better understand the factors affecting the generalizability of DL algorithms. External validation and prospective trials are necessities for the integration of AI based models into clinical practice.

Concluding remarks

A number of studies, including Chapters 3,4,6-11 of this thesis, investigated the effects of variations in a number of parameters on the reproducibility of HRFs [5,12-14]. The robust radiomics analysis workflow proposed in Chapter 2 was based on rigorous literature search and previous experiments. The validation and slight modification of the workflow presented in Chapters 6 and 7 provided a solid argument for the application of the workflow in radiomic studies including scans acquired with different imaging parameters.

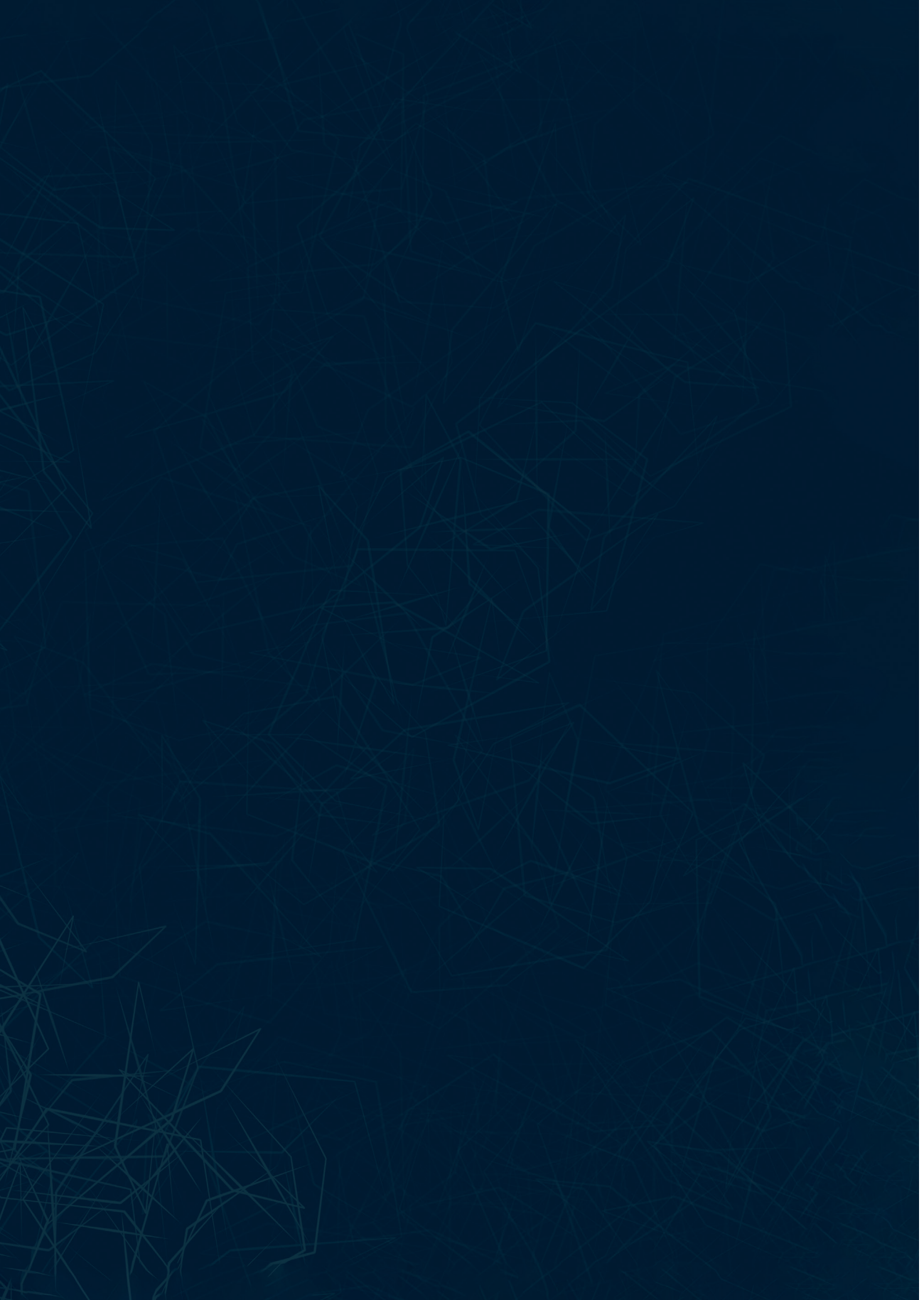
While many studies have reported on the sensitivity of HRFs to variations in imaging parameters [14-18], there has been no quantitative method to assess the reproducibility of HRFs across different imaging parameters. MPenn radiomics reproducibility score can serve as an initial tool to include/exclude scans that show low percentages of reproducible HRFs. The score was found to be robust in confirmatory analyses. The score is subject to further improvement when more data is available for analysis.

The DL based softwares presented in Chapters 12 and 13 add to the evidence showing the potential uses of DL to perform different tasks in medical image analysis. The experiments showcased the potential of DL applications for clinical decision support. Prospective validation of the developed softwares is essential to ease its translation into clinical applications. Both methods, handcrafted radiomics and DL, have great potential for prospective applications in clinical settings, given the standardization of the quantitative features and the extensive validation of the developed algorithms.

Reference

1. Lambin, P.; Leijenaar, R.T.; Deist, T.M.; Peerlings, J.; de Jong, E.E.; van Timmeren, J.; Sanduleanu, S.; Larue, R.T.; Even, A.J.; Jochems, A. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology* **2017**, *14*, 749.
2. Gillies, R.J.; Kinahan, P.E.; Hricak, H. Radiomics: images are more than pictures, they are data. *Radiology* **2015**, *278*, 563-577.
3. Rogers, W.; Thulasi Seetha, S.; Refaee, T.A.; Lieverse, R.I.; Granzier, R.W.; Ibrahim, A.; Keek, S.A.; Sanduleanu, S.; Primakov, S.P.; Beuque, M.P. Radiomics: from qualitative to quantitative imaging. *The British journal of radiology* **2020**, *93*, 20190948.
4. Morin, O.; Vallières, M.; Jochems, A.; Woodruff, H.C.; Valdes, G.; Braunstein, S.E.; Wildberger, J.E.; Villanueva-Meyer, J.E.; Kearney, V.; Yom, S.S. A deep look into the future of quantitative imaging in oncology: a statement of working principles and proposal for change. *International Journal of Radiation Oncology* Biology* Physics* **2018**.
5. Zhao, B.; Tan, Y.; Tsai, W.-Y.; Qi, J.; Xie, C.; Lu, L.; Schwartz, L.H. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Scientific reports* **2016**, *6*, 1-7.
6. Xu, Y.; Lu, L.; Sun, S.H.; Lian, W.; Yang, H.; Schwartz, L.H.; Yang, Z.-h.; Zhao, B. Effect of CT image acquisition parameters on diagnostic performance of radiomics in predicting malignancy of pulmonary nodules of different sizes. *European Radiology* **2021**, 1-11.
7. Traverso, A.; Wee, L.; Dekker, A.; Gillies, R. Repeatability and reproducibility of radiomic features: a systematic review. *International Journal of Radiation Oncology* Biology* Physics* **2018**, *102*, 1143-1158.
8. Orlhac, F.; Buvat, I. Comment on Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848. *Cancers* **2021**, *13*, 3037.
9. Ibrahim, A.; Refaee, T.; Primakov, S.; Barufaldi, B.; Acciavatti, R.J.; Granzier, R.W.Y.; Hustinx, R.; Mottaghy, F.M.; Woodruff, H.C.; Wildberger, J.E.; et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers* **2021**, *13*, 1848.
10. Walsh, S.; de Jong, E.E.; van Timmeren, J.E.; Ibrahim, A.; Compter, I.; Peerlings, J.; Sanduleanu, S.; Refaee, T.; Keek, S.; Larue, R.T. Decision support systems in oncology. *JCO clinical cancer informatics* **2019**, *3*, 1-9.
11. Ibrahim, A.; Vallières, M.; Woodruff, H.; Primakov, S.; Beheshti, M.; Keek, S.; Sanduleanu, S.; Walsh, S.; Morin, O.; Lambin, P. Radiomics Analysis for Clinical Decision Support in Nuclear Medicine. In Proceedings of the Seminars in Nuclear Medicine, 2019.
12. Balagurunathan, Y.; Kumar, V.; Gu, Y.; Kim, J.; Wang, H.; Liu, Y.; Goldgof, D.B.; Hall, L.O.; Korn, R.; Zhao, B. Test–retest reproducibility analysis of lung CT image features. *Journal of digital imaging* **2014**, *27*, 805-823.
13. Hu, P.; Wang, J.; Zhong, H.; Zhou, Z.; Shen, L.; Hu, W.; Zhang, Z. Reproducibility with repeat CT in radiomics study for rectal cancer. *Oncotarget* **2016**, *7*, 71440.

14. Baefßler, B.; Weiss, K.; Dos Santos, D.P. Robustness and reproducibility of radiomics in magnetic resonance imaging: a phantom study. *Investigative radiology* **2019**, *54*, 221-228.
15. Mackin, D.F., Xenia; Zhang, Lifei; Fried, David; Yang, Jinzhong; Taylor, Brian; Rodriguez-Rivera, Edgardo; Dodge, Cristina; Jones, Aaron Kyle; and Court, Laurence. Data From Credence Cartridge Radiomics Phantom CT Scans. *The Cancer Imaging Archive* **2017**, doi:<http://doi.org/10.7937/K9/TCIA.2017.zuzrml5b>.
16. Mahmood, U.; Apte, A.; Kanan, C.; Bates, D.D.; Corrias, G.; Manneli, L.; Oh, J.H.; Erdi, Y.E.; Nguyen, J.; Deasy, J.O. Anatomically Informed 3D Printed CT phantoms: The First Step of a Pipeline To Identify Robust Quantitative Radiomic Features. *bioRxiv* **2019**, 773879.
17. Rai, R.; Holloway, L.C.; Brink, C.; Field, M.; Christiansen, R.L.; Sun, Y.; Barton, M.B.; Liney, G.P. Multicenter evaluation of MRI-based radiomic features: A phantom study. *Medical physics* **2020**, *47*, 3054-3063.
18. Dreher, C.; Kuder, T.; König, F.; Mlynarska-Bujny, A.; Tenconi, C.; Paech, D.; Schlemmer, H.-P.; Ladd, M.; Bickelhaupt, S. Radiomics in diffusion data: a test–retest, inter-and intra-reader DWI phantom study. *Clinical Radiology* **2020**, *75*, 798. e713-798. e722.



PART VI

A large, white, serif capital letter 'A' is centered on a blue, textured, watercolor-like background. The background consists of various shades of blue, from light to dark, with a mottled, organic appearance. The letter 'A' is a classic serif font, with a triangular top and a wide base. The overall composition is simple and visually striking due to the contrast between the white letter and the vibrant blue background.

A

Appendices

Impact Paragraph
Summary
List of Publications
Acknowledgements
Curriculum Vitae

Impact Paragraph

The successful application of artificial intelligence (AI) methods across different disciplines has given a boost for the research into the application of AI in medical image analysis. The vision is that AI applications could provide fast, reliable and cost-effective tools that would aid clinicians to make personalized decisions in a significantly shorter time. Several methodologies of AI have been developed and investigated as clinical decision support systems. The focus in this thesis was on two AI methods: handcrafted radiomics and deep learning. For handcrafted radiomics, the aims were to understand the impact of variations in imaging parameters on the reproducibility of handcrafted radiomic features (HRFs), and to devise a methodology to assess the reproducibility of HRFs across different imaging parameters. Several experiments were performed to understand these effects, and to develop a novel method for assessing the reproducibility of HRFs. For deep learning, a number of potential applications were investigated, with a special attention to the explainability of developed algorithms and their comparability to current gold standards.

Scientific impacts

1. The experiments in this thesis, and the analyses codes used, are published in well-cited open access scientific journals (e.g Nature communications, Nature Scientific Reports, Cancers), which will ease the transmissibility in the academic societies.
2. Chapter 2 is an introduction to the current applications of radiomics in medical image analysis, the challenges the field face, with a proposal of a new framework that guides the development of robust radiomics signature.
3. Chapters 3 and 4 showed the difficulty of interpreting radiomics analyses results in studies analyzing scans acquired with different imaging parameters, and highlighted the need for reproducibility analyses in radiomics studies.
4. Chapter 5 reported extensively on the different harmonization methods currently used in radiomics analyses. It also highlighted the need for radiomics specific harmonization methods.
5. Chapters 6, 7 and 11 are phantom experiments that added to the understanding of how differences in imaging parameters affect the reproducibility of HRFs, and how harmonization methods like image resampling and ComBat harmonization perform in different scenarios.

Impact Paragraph

6. Chapters 8, 9 and 10 are patient scans experiments that highlighted the effects of differences in a number of parameters (test-retest, inter-observer, and different imaging phases) on the reproducibility of HRFs.
7. Chapter 11 presented a novel quantitative score that could be used in future radiomics studies to assess the reproducibility of HRFs across the scans available for analysis.
8. Chapters 12 and 13 showcase the potential application of DL algorithms for different clinical endpoints, with one of the softwares being CE marked.

Social impacts

1. The standardization of handcrafted radiomic features will ease the generalization of developed radiomic signatures across different institutions.
2. Developing generalizable and robust radiomic signatures will ease the integration of these signatures in clinical decision support systems.
3. Radiomics has the potential to improve patient care by guiding personalized management rather than the fit-for-all approach, which is also done in less invasive manners.
4. Radiomics can provide a cost-effective means that would reduce health expenditure and improve public resources management.

Summary

Recent decades witnessed an exponential growth in the number of studies investigating the potential applications of artificial intelligence (AI) in medical image analysis. Handcrafted radiomics is one of the methods that employ AI methods in medical image analysis. Handcrafted radiomic features (HRFs) are quantitative features extracted from medical images by applying handcrafted formulas on the array of values representing a given medical image. The general hypothesis is that HRFs can decode biologic characteristics, and could be used potentially to personalize patients care. In addition, handcrafted radiomics can become an alternative to current gold standard diagnostic methods, as with proper development, it can be a non-invasive and time-saving clinical tool. HRFs have been reported to correlate with different clinical endpoints, such as classification of lesions on medical images, predicting response to therapy, and survival.

Despite the reported high potential of handcrafted radiomics, a number of limitations that hinders the clinical integration of radiomics signatures have been identified. HRFs have been reported to be sensitive to inter-reader variability, test-retest, and to variations in imaging parameters, in addition to the need for large datasets. In this thesis, we performed experiments to validate these hypotheses. We confirmed that HRFs are sensitive to the above mentioned variations, using phantom and patients reproducibility studies. We further hypothesized that different harmonization methods will have different effects on different HRFs. We performed experiments to assess the impacts of different harmonization methods, mainly image resampling and ComBat harmonization. Lastly, we hypothesized that a quantitative tool can be developed based on the differences in imaging parameters. Our novel MPenn radiomics reproducibility score was developed using a large number of scenarios of variations in imaging parameters, and has shown robustness and high performances in assessing the percentage of reproducible HRFs across scans acquired differently. The score can be utilized in future radiomics studies to evaluate the agreement in HRFs, if the scans to be analyzed are acquired differently. The score would also help interpreting the results of radiomics analyses.

Additionally, we performed a number of experiments to assess potential applications of deep learning (DL), the other AI method investigated in this thesis, in medical image analysis. Classification of bone scintigraphies and the automated detection and segmentation of non-small cell lung carcinoma on CT scans are the two tasks investigated in this thesis. For each of the tasks, multicenter data was collected, and a relatively large number of medical images were used to train the DL algorithms. A partition of the collected datasets was kept for external validation of developed algorithms. In addition, *in silico* trials to assess the performance of developed algorithms were designed for each of the tasks investigated. Our results showcased the potential of DL algorithms to be

Summary

used as clinical decision support tools, with one of the developed algorithms receiving CE mark.

In conclusion, this thesis has confirmed a number hypothesis regarding the applications of handcrafted radiomics and deep learning in medical image analysis. For handcrafted radiomics, we proposed and assessed a workflow for robust handcrafted radiomics analyses that will help developing generalizable radiomics signatures; and developed a novel quantitative method to assess the reproducibility of HRFs across scans acquired differently. For DL, we assessed and showcased the potential of automated algorithms to aid clinical decision making. We developed a DL algorithm for three different tasks, which showed high performance and prospective for clinical applications.

List of Publications

1. Stratakis, N., Gielen, M., Margetaki, K., Godschalk, R.W., Van der Wurff, I., Rouschop, S., **Ibrahim, A.**, Antoniou, E., Chatzi, L., de Groot, R.H.M. and Zeegers, M.P., 2017. Polyunsaturated fatty acid levels at birth and child-to-adult growth: Results from the MEFAB cohort. *Prostaglandins, Leukotrienes and Essential Fatty Acids*, 126, pp.72-78.
2. **Ibrahim, A.**, Abdalla, S.M., Jafer, M., Abdelgadir, J. and De Vries, N., 2019. Child labor and health: a systematic literature review of the impacts of child labor on child's health in low-and middle-income countries. *Journal of Public Health*, 41(1), pp.18-26.
3. Walsh, S., de Jong, E.E., van Timmeren, J.E., **Ibrahim, A.**, Compter, I., Peerlings, J., Sanduleanu, S., Refaee, T., Keek, S., Larue, R.T. and van Wijk, Y., 2019. Decision support systems in oncology. *JCO clinical cancer informatics*, 3, pp.1-9.
4. Peerlings, J., Woodruff, H.C., Winfield, J.M., **Ibrahim, A.**, Van Beers, B.E., Heerschap, A., Jackson, A., Wildberger, J.E., Mottaghy, F.M., DeSouza, N.M. and Lambin, P., 2019. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Scientific reports*, 9(1), pp.1-10.
5. **Ibrahim, A.**, Vallieres, M., Woodruff, H., Primakov, S., Beheshti, M., Keek, S., Sanduleanu, S., Walsh, S., Morin, O., Lambin, P. and Hustinx, R., 2019, September. Radiomics analysis for clinical decision support in nuclear medicine. In *Seminars in nuclear medicine* (Vol. 49, No. 5, pp. 438-449). WB Saunders.
6. Sanduleanu, S., Wiel, A., Lieveise, R.I., Marcus, D., **Ibrahim, A.**, Primakov, S., Wu, G., Theys, J., Yaromina, A., Dubois, L.J. and Lambin, P., 2020. Hypoxia PET Imaging with [18F]-HX4—A Promising Next-Generation Tracer. *Cancers*, 12(5), p.1322.
7. Kalmet, P.H., Sanduleanu, S., Primakov, S., Wu, G., Jochems, A., Refaee, T., **Ibrahim, A.**, Hulst, L.V., Lambin, P. and Poeze, M., 2020. Deep learning in fracture detection: a narrative review. *Acta orthopaedica*, 91(2), pp.215-220.
8. Refaee, T., Wu, G., **Ibrahim, A.**, Halilaj, I., Leijenaar, R.T.H., Rogers, W., Gietema, H.A., Hendriks, L.E.L., Lambin, P. and Woodruff, H.C., 2020. The Emerging Role of Radiomics in COPD and Lung Cancer. *Respiration; International Review of Thoracic Diseases*, 99(2), pp.99-107.
9. Wu, G., Woodruff, H.C., Shen, J., Refaee, T., Sanduleanu, S., **Ibrahim, A.**, Leijenaar, R.T., Wang, R., Xiong, J., Bian, J. and Wu, J., 2020. Diagnosis of invasive lung adenocarcinoma based on chest CT radiomic features of part-solid pulmonary nodules: a multicenter study. *Radiology*, 297(2), pp.451-458.
10. Rogers, W., Thulasi Seetha, S., Refaee, T.A., Lieveise, R.I., Granzier, R.W., **Ibrahim, A.**, Keek, S.A., Sanduleanu, S., Primakov, S.P., Beuque, M.P. and Marcus, D., 2020. Radiomics: from qualitative to quantitative imaging. *The*

- British journal of radiology*, 93(1108), p.20190948.
11. Granzier, R.W.Y., Verbakel, N.M.H.*, **Ibrahim, A.***, van Timmeren, J.E., van Nijnatten, T.J.A., Leijenaar, R.T.H., Lobbes, M.B.I., Smidt, M.L. and Woodruff, H.C., 2020. MRI-based radiomics in breast cancer: feature robustness with respect to inter-observer segmentation variability. *Scientific reports*, 10(1), pp.1-11.
 12. **Ibrahim, A.**, Primakov, S., Beuque, M., Woodruff, H.C., Halilaj, I., Wu, G., Refaee, T., Granzier, R., Widaatalla, Y., Hustinx, R. and Mottaghy, F.M., 2021. Radiomics for precision medicine: Current challenges, future prospects, and the proposal of a new framework. *Methods*, 188, pp.20-29.
 13. Granzier, R.W., **Ibrahim, A.**, Primakov, S.P., Samiei, S., van Nijnatten, T.J., de Boer, M., Heuts, E.M., Hulsmans, F.J., Chatterjee, A., Lambin, P. and Lobbes, M.B., 2021. MRI-Based Radiomics Analysis for the Pretreatment Prediction of Pathologic Complete Tumor Response to Neoadjuvant Systemic Therapy in Breast Cancer Patients: A Multicenter Study. *Cancers*, 13(10), p.2447.
 14. Samiei, S.*, Granzier, R.W.*, **Ibrahim, A.**, Primakov, S., Lobbes, M.B., Beets-Tan, R.G., van Nijnatten, T.J., Engelen, S.M., Woodruff, H.C. and Smidt, M.L., 2021. Dedicated Axillary MRI-Based Radiomics Analysis for the Prediction of Axillary Lymph Node Metastasis in Breast Cancer. *Cancers*, 13(4), p.757.
 15. Wu, G., Jochems, A., **Ibrahim, A.**, Yan, C., Sanduleanu, S., Woodruff, H.C. and Lambin, P., 2021. Structural and functional radiomics for lung cancer. *European Journal of Nuclear Medicine and Molecular Imaging*, pp.1-14.
 16. Mali, S.A., **Ibrahim, A.**, Woodruff, H.C., Andrearczyk, V., Müller, H., Primakov, S., Salahuddin, Z., Chatterjee, A. and Lambin, P., 2021. Making Radiomics More Reproducible across Scanner and Imaging Protocol Variations: A Review of Harmonization Methods. *Journal of Personalized Medicine*, 11(9), p.842.
 17. **Ibrahim, A.**, Refaee, T., Leijenaar, R.T., Primakov, S., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Maidment, A.D. and Lambin, P., 2021. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *Plos one*, 16(5), p.e0251147.
 18. **Ibrahim, A.**, Refaee, T., Leijenaar, R.T., Primakov, S., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Maidment, A.D. and Lambin, P., 2021. The application of a workflow integrating the variable reproducibility and harmonizability of radiomic features on a phantom dataset. *Plos one*, 16(5), p.e0251147.
 19. **Ibrahim, A.**, Refaee, T., Primakov, S., Barufaldi, B., Acciavatti, R.J., Granzier, R.W.Y., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Wildberger, J.E., Maidment, A.D.A, and Lambin, P., 2021. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features' Stability with and without ComBat Harmonization. *Cancers*, 13(8).
 20. **Ibrahim, A.**, Refaee, T., Primakov, S., Barufaldi, B., Acciavatti, R.J., Granzier, R.W., Hustinx, R., Mottaghy, F.M., Woodruff, H.C., Wildberger, J.E., Lambin,

- P. and Maidment, A.D., 2021. Reply to Orhac, F.; Buvat, I. Comment on “Ibrahim et al. The Effects of In-Plane Spatial Resolution on CT-Based Radiomic Features’ Stability with and without ComBat Harmonization. *Cancers* 2021, 13, 1848”. *Cancers*, 13(12), p.3080.
21. Granzier, R., **Ibrahim, A.**, Primakov, S., Keek, S., Halilaj, I., Zwanenburg, A., Engelen, S., Lobbes, M., Lambin, P., Woodruff, H. and Smidt, M. (2021), Test–Retest Data for the Assessment of Breast MRI Radiomic Feature Repeatability. *J Magn Reson Imaging*. <https://doi.org/10.1002/jmri.28027>
 22. **Ibrahim, A.**, Widaatalla, Y., Primakov, S., Miclea, R.L., Öcal, O., Fabritius, M.P., Ingrisich, M., Ricke, J., Hustinx, R., Mottaghy, F.M. and Woodruff, H.C., 2021. Reproducibility of CT-Based Hepatocellular Carcinoma Radiomic Features across Different Contrast Imaging Phases: A Proof of Concept on SORAMIC Trial Data. *Cancers*, 13(18), p.4638.
 23. **Ibrahim, A.**, Barufaldi, B., Refaee, T., Filho, T.M.S., Acciavatti, R.J., Salahuddin, Z., Hustinx, R., Mottaghy, F.M., Maidment, A.D.A, and Lambin, P. MaasPenn radiomics reproducibility score: a novel quantitative measure for evaluating the reproducibility of CT-based handcrafted radiomic features. **Submitted to JAMA.**
 24. **Ibrahim, A.**, Vaidyanathan, A., Primakov, S., Belmans, F., Bottari, F., Refaee, T., Lovinfosse, P., Jadoul, A., Derwael, C., Hertel, F., Woodruff, H.C, Zacho, H.D, Walsh, S., Vos, W., Occhipinti, M., Hanin, F.X, Lambin, P., Mottaghy, F.M, Hustinx, R. Deep learning based identification of bone scintigraphies containing metastatic bone disease foci. ***Under review in EJNMMI.***
 25. Primakov, S.P, **Ibrahim, A.**, van Timmeren, J.E, Wu, G., Keek, S.A., Beuque, M., Granzier, R.W.Y, Lavrova, E., Scrivener, M., Sanduleanu, S., Kayan, E., Halilaj, I., Lenaers, A., Wu, J., Monshouwer, R., Geets, X., Gietema, H.A, Hendriks, L.E.L, Morin, O., Jochems, A., Woodruff, H.C, Lambin, P. Validated fully automated detection and segmentation of non-small cell lung cancer on computed tomography images. ***Accepted with minor revisions in Nature Communications.***

Acknowledgements

First and foremost, all praise to Allah for the endurance to perform and carry on this work. As I started to write this last part of my thesis, I had a trip down memory lane of the past 42 months that I spent doing my research and finalizing this thesis. It has been an eventful chapter of life, with all its ups and downs, but I am proud to have gone through this journey. There are a lot of people without whose presence and support, this achievement would have been much harder to attain. Henceforth, although thanking them is not enough, it is a necessity.

To my promoter Professor Philippe Lambin who has been a continuous source of support and inspiration. I was lucky to have joined his group as a research assistant, following which I was accepted to do a PhD trajectory under his supervision. Prof. Lambin has continuously supported my research projects and provided motivation and uncountable chances to extend my knowledge and networks. Nevertheless, the continued guidance and academic input, as well as professional conduct, have paved the way for me to develop as a researcher. Thank you sincerely for all the efforts, opportunities and guidance that you provided over the years, which ultimately resulted in the completion of this work.

To my promoters Professors Roland Hustinx and Felix Mottaghy who were the clinical motivators behind my research. Your guidance and critical, yet constructive and friendly inputs that focused on addressing unmet clinical needs, have significantly improved its quality. Furthermore, I was always motivated to do more because of your continued motivation and support on all levels, academic and non-academic. You have played a significant role every step along the way, and for that I am eternally grateful.

My co-promoter Dr. Andrew Maidment, since we started to work together, I was fascinated by the amount of hands-on knowledge and insight into the physics of medical imaging. Throughout my experiments, Dr. Maidment has been providing extensive supervision, that I allowed to me to understand different concepts in medical physics to design my experiments. Furthermore, the smoothness, please and ease of communication and feedbacks have led to fruitful experiments and novel approaches. I was pleased to have visited the University of Pennsylvania, and work in one office with you. Thank you for all the time and guidance you have provided, it would have been much more difficult in your absence.

I would also like to thank Prof. Joachim Wildberger, who walked me baby-steps in my first endeavors to learn about the physics of computed tomography, and provided me with the opportunity to perform experiments. Your input has been critical in shaping my research ideas and plans, and has improved the quality of our joint manuscripts.

In addition, the gratitude is extended to Dr. Olivier Morin, who organized collaboration projects in the University of California-San Francisco, and provided the unique chance of working on a different PACS, and collecting a great number of images to perform different experiments. I would also like to deeply thank you for the great advices on my first paper. They have significantly improved my writing skills, and made the task much easier.

To Dr. Razvan Miclea, the radiologist who gave the idea for my first research work, and trained me to identify and segment different types of images and lesions. Your role has been very significant for the progress of all my PhD projects, but also for my future practice. I am very thankful for all the time and advices you gave during the past years, and I hope our collaboration will still thrive.

To the kind hearted that always made us feel home and welcomed, Simone. Thank you for taking care of our well-being, as well as our growth as researchers. Most importantly, you gave us the feeling that we are family. I have the deepest gratitude for you, and I wish you and your family all the best to come.

To my colleagues and office members Jurgen, Ralph, Janita, Guangyao, Zohaib, Lisa, Xian, Yi, Shruti, Relinde, Fadila, Floor, Rian, Sebastiaan, and Henry, with whom I shared not only the working space, but also great deal of joyful memories. Thank you for making my PhD journey a lot more pleasant.

To my office friends Sergey, the man behind everything. It was a great pleasure getting to know you and to work with you. Mallorca is always going to be a trip that I treasure; Simon, the first person I got to work with in the lab. It was always nice to have those discussions about research and life, and all the screaming on our online gaming nights; the flower of the office Manon, who since joined our office, we felt the sophisticated touch. We shared a house for a while, and had unforgettable times in Paris. It was a pleasure getting to know you and work with you. To my dear Iva, the nona. Thank you for all the good times we had together. It was always good to just show up at your door, have a coffee or some food together. You have made Maastricht feel more like home with your presence. To my partner in crime, Renee, you have been one of the people I worked the most with in my last years, and I can only say it was joyful to work with you. We shared all different types of moments and made memories for life. Having you along this journey has been nothing short of entertaining, and I wish you all a life full of success!

To my family, who has provided unconditional support in all the chapters of my life. My mother, Maha, the beacon of love, hope and inspiration. My words and actions of gratitude will never be enough to thank you for the role you have been playing all along.

Acknowledgements

I could not imagine what I would have become if not for you. My father, Khalil, who has always stood by my side, and trusted my choices in life. Your care and passion have been a light that never faded. To my elder sister, Walaa, and my heart-melting nephews, thank you for always encouraging me to do my best, your continuous check-ups on how things are going, and our video calls that always made me feel home. To my only brother and best friend, Mohammed, the benzene to my car. I cannot imagine how boring and unchallenging my life would have been without you. You have always had faith in me, took uncountable number of blames in my place, covered for me when I needed, and most importantly, you have always been my gaming partner who always complain about how bad am I compared to him. Today, I would like to tell you that all academic degrees are “wiped out”! To my lovely twin sisters, Sanaa and Doaa, who have brought only joy to our lives since the day they came to this world. Thank you for all your love and support that always gave me the strength to keep going. I am telling you proudly that I made it. To my aunts and uncles, Salma, Suhair, Sana, Durria, Yasir, Hisham, Nadir, and my beloved late uncle Jalal, my cousins Nada, Khalid, Ayman and Hatim, the people that supported me in all the possible ways to reach the goals I set. You are the best family I could ever ask for!

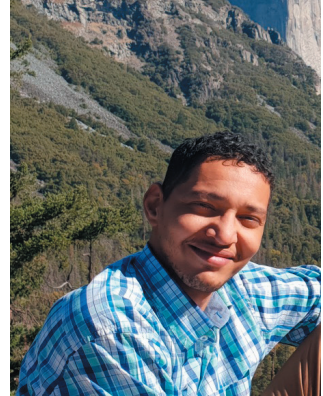
A special dedication to the family that I was lucky to find in the Netherlands. To Turkey, I cannot start to describe how grateful I am to have met you here without writing another book, so I will keep it short. As the saying goes “brother from another mother”; you have literally been like a brother. You have always been there for whatever you could help with from introducing me to the group and introducing me to the world of coding, to daily life matters not related to work. Thank you for everything, and I hope we will stay connected. To Yousif, the life-long fella, and the brother I travelled the world with. I still remember when we were sitting at your place in Khartoum, planning our “escape” to new lives more than 10 years ago. You have been the well to my secrets, the cheerleader to my games, and my backbone when I needed support. We know each other for too long now to say goodbye, but I’ll say till we live in the same city again ;). To Faisal, another brother from another mother. Thank you for the support that you gave through the years. It was always fun to have you and Ghalib, eat together and have our long chatting nights. Saudi Arabia is one of my destinations now because of you guys! I hope we will always meet in good times. To Nikos, the first non-arabic speaking friend I made in the Netherlands. You have certainly played a major role in my development as a researcher that I will always be grateful for, as you were the 1st supervisor I ever had and have learnt a great deal from you. We have also shared the same house for a while; we cooked and chatted every night, and had great times. Thank you for always supporting me, even though we are thousands of miles apart. To my vecchia Georgia, the best housemate a person could ask for. I have certainly enjoyed your company, the intelligent conversations, and the little fuses we used to make. You have motivated me to

do different things, and always had that positive energy to make everyone smile. Thank you as well for introducing me to the amazing greek culture, and the Cretan ways ;) You and Creta will always be special to me. To Jacopo, my 1st and most crazy housemate! You have always had you sarcastic, yet philosophic view that made the worst situations bearable. I will always remember how you glorify Italian food and all the good recipes from your grandmother. We shared the same roof long enough to become friends for life, thank you for always supporting me; ti voglio bene!

Last but not least, to the family that I chose for ages, and their support was a main pillar in all my achievements to date: Yeddi, Al7albi, Marwan, Aziz, Number one, Ashwal, Fatta7, Eltaib Amir, Altikaina and Khalid Abdelsadig. Even though we are thousands of miles apart, and everyone is busy with their own life, you have always found a way to keep in touch, provide all kinds of support, and lift me up when I am down. I could not have done it without you all, and I will always strive to make you proud of me, as much as I am proud of you.

Curriculum Vitae

Abdalla Ibrahim was born on the 7th of January 1991 in Medani, Sudan. Having scored among the top three hundred applicants in University entrance exams, he studied Medicine at the University of Khartoum, the most prestigious in Sudan. Following graduation in 2014, he worked shortly in Sudanese hospitals, before moving to the Netherlands to develop his research skills.



Abdalla studied global health at Maastricht University, and obtained his MSc degree in 2015. After that, he followed research internships in different fields, including epidemiology and cell cultures, before joining Prof. Lambins' research group as a research assistant.

He showed great interest in the field of radiomics, and developed the necessary skills in a short time. He started his PhD trajectory in July 2018 after obtaining Liege-Maastricht imaging valley grant, a double PhD degree awarded by Universities of Liege and Maastricht.

Awards

2018. Liege-Maastricht Imaging Valley Research Grant, awarded by the Universities of Liege and Maastricht.

2020. GROW school Travel Grant, awarded by GROW, the School for Oncology and Developmental Biology at the Maastricht University Medical Centre (MUMC+), The Netherlands.